(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
31 May 2001 (31.05.2001)

PCT

(10) International Publication Number
WO 01/39098 A2

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/IL00/00779

(22) International Filing Date:
22 November 2000 (22.11.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/166,744 22 November 1999 (22.11.1999) US
60/209,806 7 June 2000 (07.06.2000) US

(71) Applicant *(for all designated States except US)*: **YISSUM RESEARCH DEVELOPMENT COMPANY OF THE HEBREW UNIVERSITY OF JERUSALEM** [IL/IL]; 46 Jabotinsky Street, P.O. Box 4279, 92182 Jerusalem (IL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **GOLDBLUM, Amiram** [IL/IL]; Shimshon 20, 93501 Jerusalem (IL). **GLICK, Meir** [IL/IL]; Sheshet Ha'Yamim 2, 59503 Bat Yam (IL).

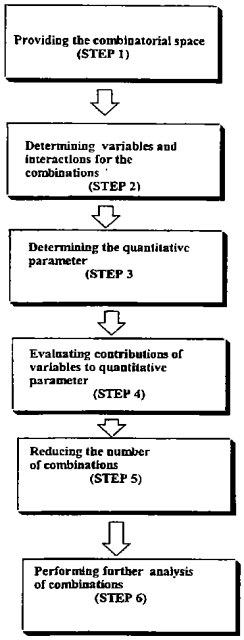
(74) Agent: **BODNER, Marc**; Plinner, Bodner, Brass, Beit Agish-Ravad, 13 Noach Mozes Street, 67442 Tel Aviv (IL).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DE (utility model), DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR SEARCHING A COMBINATORIAL SPACE



(57) Abstract: A system and method for searching through combinatorial space, without a combinatorial explosion. The search is performed for various combinations of basic elements, according to at least one desired property of the combination, which is translatable into a quantitative measurement of the success of the search. Since the number of variables and hence the number of combinations may be very large, preferably samples of combinations are examined. Those elements of the combinations which have consistent maximization and/or promotion of the quantitative measurement are then kept, while the other elements are dropped. This process is then repeated until some minimum number of combinations is found, which could then optionally be further evaluated according to a similar parameter and/or some other parameter or characteristic.

WO 01/39098 A2

IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM AND METHOD FOR SEARCHING A COMBINATORIAL SPACE

FIELD OF THE INVENTION

The present invention discloses a system and method for searching through a combinatorial space, and in particular, to such a system and method in which one or more combinations of basic elements having a desired property can be located in the combinatorial space. The desired property should have a numerical basis, or at the very least should be translatable into some type of numerical measurement and/or equivalent. The present invention enables the combinatorial space to be searched rapidly and efficiently according to the property, without a combinatorial explosion. The present invention accomplishes these tasks by examining each value of a basic element at least once, and preferably a plurality of times, during the search process. Therefore, each value can be said to be searched in an exhaustive search, yet not every combination of the values for the basic elements needs to be exhaustively searched. Thus, the present invention combines the efficiency of non-exhaustive, stochastic search processes with the efficacy of exhaustive searches.

BACKGROUND OF THE INVENTION

This section has been divided into a number of different sub-sections for ease of explanation. Briefly, the first section describes the general problem of combinatorial spaces, and searching within these spaces. The subsequent sections describe previous attempted solutions at solving a number of different biological problems, which are examples of inadequacy of background art solutions to handle combinatorial search spaces with regard to biological problems. These sections include placement of polar protons for biological molecules such as proteins; placement of side chains for amino acids in proteins; and prediction of loop structures in proteins.

Combinatorial Spaces

A combinatorial space is defined as having multiple combinations of basic elements. These combinations may differ according to the values of different types of these elements, the structure of the resultant combination of elements, or may be produced as a result of both factors. At a more basic level, each combination may be considered to consist of *variables*, each of which may assume more than a single value. For the purposes of the present description, each variable preferably assumes one value of a set of discrete values, although

alternatively, each variable may assume one value out of a range of continuous values or out of a function, for example.

Combinatorial spaces often occur in biology, as many elementary biological materials are themselves produced through combinations of relatively basic building blocks, yet are highly complex in their resultant structure and/or function. Examples of combinatorial spaces include, but are not limited to, proteins, which are produced from combinations of amino acids as building blocks and eventually fold into a spatial structure known as a "tertiary structure", which is one set of *values* in the combinatorial space. A search for this single structure through such a combinatorial space may also be termed a "combinatorial search". However, the folded protein in its biological environment is not fixed in a single "tertiary structure" but may exist in many conformational substates that are in equilibrium. Thus, searching through combinatorial space should preferably find more than a single solution.

Searching through combinatorial space is a difficult problem since, despite the apparent simplicity of these different types of building blocks, the huge number and complexity of the resultant combinations make an exhaustive search of the combinatorial space beyond the scope of state of the art computers. For example, a simple, small protein having a relatively short amino acid sequence still has a huge number of different potential structures, yet has only one or a few actual stable structures. With regard to protein structure, the problem is composed of several sub problems such as positioning side chains of the amino acids, polar protons (for determining hydrogen bonding within the protein and optionally also between the protein and another molecule), and also the location and type of larger structures within the protein such as loops. Thus, searching through these types of combinatorial spaces, particularly for biological problems, has typically proved to be resistant to modeling and prediction by various computational approaches.

Many attempts have been made to handle the problem of searches through combinatorial space. Generally, these attempts have included directed searches, in which the search strategy is guided toward the desired solution; transformation of the larger search space into a smaller space, for example by refinement of the combinatorial space; altering the representation of the space; and altering the criteria for locating a potential solution to be less deterministic. However, none of these solutions has proven to be suitable for biological problems such as predicting protein structure. Directed searches are not useful for these problems, since many "dead ends" in protein structure are possible, and a definitive set of

rules for protein folding is not known. Similarly, the search space cannot currently be reduced since these rules for protein structure are not known. Alteration of the representation of the combinatorial space is not possible, since protein structure has basic, fixed building blocks, which are amino acids. Finally, locating less deterministic solutions has not proven useful for general prediction of protein structure, since the apparent "rules" which may be derived for folding one protein are themselves currently specific for that protein, and have not been generalized. There is no efficient search strategy that can detect the population of the optimal set of *values* in combinatorial space.

First Biological Problem: Polar Protons

These general attempted solutions may be further described in terms of specific solutions for particular biological problems. For example, the positions of polar protons (hydrogens) are crucial for determining hydrogen bonding relationships and specificities within biologically important molecules such as proteins and DNA molecules, as well as for determining such hydrogen bonding between these molecules and other molecules. Inclusion of all hydrogen atoms in protein and nucleic acid models is necessary for a more accurate representation of biological systems during energy minimizations, molecular dynamics simulations, and for understanding molecular recognition (Jones et al., J Mol Biol 1995;24:43-53). Polar hydrogens play a critical role in determining secondary structure and protein packing, and the exact placement and the ensuing formation of hydrogen bonds is extremely important for energy evaluation. One misplaced polar hydrogen in an active site has been shown to drastically change the substrate conformation during molecular dynamics simulations (Bass et al., Proteins 1992; 12:266-277).

At present, X-ray crystallography is still the main source for acquiring high resolution data of biomolecules. However, it is efficient for locating heavy atoms while the proton positions remain undetermined yet. Neutron diffraction studies can locate the protons but to date, only a few combined X-rays/Neutron diffraction studies have been deposited in the protein data bank (PDB).

A few computer based methods have been proposed to place polar hydrogens in a protein structure. The first, common to most molecular modeling software packages, places hydrogens in a non specific manner, and then one may optimize the structure by energy minimization algorithms that suffer from the multiple minima problem: they do not take into account the alternative positions of flexible polar hydrogens, many of which could form

hydrogen bonds.

A second method suggested by Brunger and Karplus (Proteins 1988; 4:148-156), employs a search for the local minimum conformation of each polar proton in its turn, by torsion angle rotations. Then an iterative process continues until convergence is reached. This method does not consider the effects of neighboring rotatable hydrogens, and therefore would be accurate only for systems in which such close contacts between hydrogens are absent.

The third method suggested by Bass et al. (Proteins 1992; 12:266-277) is based on dividing the system into networks of interacting hydrogen bond donors and acceptors. The algorithm tries to maximize the number of hydrogen bonds that can be formed in each network and to minimize the total distance between donors and acceptors. Because each network is rigorously examined for the best possible set of hydrogen bonds, the number of comparisons (between options) scales with the factorial of the number of elements in the network-a fact that limits the calculation to small networks (Bass et al., Proteins 1992; 12:266-277). No energy evaluations are employed to choose the best structure. As a result, the output might contain high energy interactions between the located hydrogens and their environment.

Richardson et al. (J. Mol. Biol. 1999; 285:1711-1733) and Word et al., (J. Mol. Biol. 1999; 285:1735-1747) have recently extended the "network" approach by taking into account the Asn/Gln "flips", the protonation state of histidine rings, and a simple model of water interactions. For rotatable protons, a set of local H-bonds is optimized for distances and Van der Waals overlaps. Unfortunately, none of these attempted solutions, for the problem of placement of polar protons in biological molecules such as proteins, can be generalized to even a class of such molecules, accurate within the class, and efficient for execution.

Second Biological Problem: Placement of Amino Acid Side Chains

Another example of a problem for searching through combinatorial space is the placement of the side chains of amino acids. Even though this problem is itself solved through combinatorial space, it is only one part of the general problem of protein structure prediction. However, this problem has so far proved to be intractable to currently available methods for attempting to predict the locations of these side chains.

Accurate placement of protein side chains is essential for both theoretical and experimental purposes. On the theoretical side, it is a sub-problem in *de novo* protein

structure prediction. It is imperative for structure based drug design (Defay & Cohen, Proteins 1995; 23:431-445), for inverse folding and threading algorithms (Bahar & Jernigan, J. Mol. Biol. 1997; 266: 195-214), for understanding the folding process and structural stability (Zhukov et al., Protein Sci. 2000; 9: 273-279), ab-initio predictions of protein tertiary structure (Huang et al., Proteins 1998; 33:204-217) and for homology-based modeling (Blundell et al., Nature 1987; 326: 347-352). From the X-ray crystallographer's point of view, it could speed the placement of side chains using the electron density maps of the main chain prior to refinement calculations. The main limitation is the large amount of possible conformations that each side chain may assume (Lee & Subbiah, J. Mol. Biol. 1991; 217: 373-388). An exhaustive search of all possible protein conformations is beyond the scope of state of the art computers.

X-ray crystallography usually supplies a single structure characterized by an R-factor. A crystal structure reflects the biomolecule in the highly ordered crystal lattice, as opposed to the more physiologically relevant solution environment of a NMR structure. The former might be biased toward specific conformational substates in the crystal, which may not be among the ensemble of conformations in solution (Brunger, Nat. Struct. Biol. 1997; 4 suppl: 862-865). Observation of alternate rotamers is beyond the detection limits of conventional X-ray crystallographic techniques, except at the very highest resolution. At least 10% of all side chains in proteins adopt multiple, discrete conformations in carefully refined crystal structures (Smith et al., Biochemistry, 1986; 25: 5018-5027). MacArthur & Thornton (Acta. Cryst. D Biol. Cryst. 1999; D55: 994-1004) found a significant and unexpected correlation between χ_1 mean values and resolution mainly for small flexible side chains. All the data support the hypothesis that this observation reflects local conformational flexibility and disorder, which at low resolution might be interpreted as a single distorted conformer. The results of all these investigations point to a dynamic, rather than static, picture of protein structure and to the need of extracting this dynamic information from NMR ensembles to gain a more detailed understanding of protein function (Philippopoulos & Lim, Proteins 1999; 36: 87-110). Protein function and molecular recognition depend on structural plasticity (Garcia et al., Science. 1998; 279: 1166-1172) and conformational flexibility of a receptor protein is one of the major factors affecting ligand docking (Desmet et al., FASEB J. 1997; 11: 164-172). However, accurate computer location of protein side chains is a complicated task, due to the large number of minimum energy conformers on the potential energy surface, even with a rigid backbone. Conventional methods for side chain addition usually result in a

single structure of the protein, which is then compared to the X-ray structure, if available. The conformational space is disregarded. NMR studies of many proteins have been conducted in recent years and many conformations for each protein are suggested (Schneider et al., J. Mol. Biol. 1999; 285: 727-740). It is however not clear if such conformations represent alternative solutions of the distance restrictions that emerge from the 2D and 3D coupling maps or, they are real conformations that may contribute to the overall population at equilibrium. Classical molecular dynamics (MD) methodology is the technique of choice for simulating biomolecules. With current technology, MD simulations of systems consisting of tens of thousands of atoms for a few nanoseconds are becoming more common (Sagui & Darden, Ann. Rev. Biophys. Biomol. Struct. 1999; 28: 155-179). However, relevant time scales for biomolecular function range from nanoseconds to more than seconds. The time required to reach an equilibrium between different conformers of a protein by MD is prohibitive for such simulations, and we may acquire only a glimpse of the protein's behavior in its surrounding.

Current strategies for side chain addition differ in three categories. The first is the conformational space of each side chain. In continuous space methods (Eisenmenger. J. Mol. Biol. 1993; 231: 849-860; Roitberg. & Elber, J. Chem. Phys. 1991; 95: 9277-9287), any side-chain torsion angle may be sampled. Discrete space methods are based on the assumption that side-chains exist in energetically preferred conformations called rotamers, which are local minima conformers that have been sampled by statistical analysis of known structures (Chandrasekaran & Ramachandran, Int. J. Protein Res. 1970; 2: 223-233; Sasisekharan & Ponnuswamy, Biopolymers 1970; 9; 1249-1256; Sasisekharan & Ponnuswamy, Biopolymers 1971; 10: 583-592; Ponder & Richards J. Mol. Biol. 1987; 193: 775-791; Gelin & Karplus, Biochemistry 1979; 18: 1256-1268; Dunbrack & Karplus, Nat. Struct. Biol. 1994; 1: 334-340). Discrete space methods can not predict conformations that are not present in the rotamer database. But large rotamer databases which contain very rare conformations do not necessarily yield better predictions than smaller databases (Holm & Sander, Proteins 1992;14: 213-223; Laughton, J. Mol. Biol. 1994; 235: 1088-1097; Tanimura et al., Protein Sci. 1994; 3: 2358-2365; Vasquez, Biopolymers 1995; 36: 53-70). Databases can also be classified into backbone dependent and backbone independent. The first are based on a relationship between the side-chain conformation and the local backbone conformation, while the latter are not.

The second category is the cost function for evaluating solutions. Energy based

methods rely on non-bonded terms (Laughton, J. Mol. Biol. 1994; 235: 1088-1097; Vasquez, Biopolymers 1995; 36: 53-70; Wilson et al., J. Mol. Biol. 1993; 229: 996-1006; Vasquez, Curr. Opin. Struct. Biol. 1996; 6: 217-221). The assumption is that the lower the energy, the more accurate the prediction.

Knowledge based methods were also proposed: Sutcliffe et al. (Protein Eng. 1987; 1: 385-392) suggested a procedure for building side chains using spatial information from side chains in topologically equivalent positions -as far as such a correlation may be observed- and most probable conformations of the side chains in the respective secondary structure type. Sali & Blundell (J. Mol. Biol. 1993; 234: 779-815) described a comparative protein modelling method designed to find the most probable structure for a sequence, given its alignment with related structures. Bower et al. (J. Mol. Biol. 1997; 267: 1268-1282) located residues in their most favorable backbone-dependent rotamers and systematically resolved the conflicts that arise from that structure.

The third category is the search strategy. Examples for search strategies being employed are various. Metropolis Monte Carlo methods (Holm & Sander, Proteins 1992; 14: 213-223), Gibbs sampling Monte Carlo (Vasquez, Biopolymers 1995; 36: 53-70), Neural networks (Hwang & Liao, Protein Eng. 1995; 8: 363-370), Genetic Algorithms (Tuffery et al., J. Biomol. Struct. Dynam. 1991; 8: 1267-1289; Tuffery et al., J. Comput. Chem 1993; 14: 790-798), Simulated Annealing (Lee & Subbiah, J. Mol. Biol. 1991; 217: 273-288), Mean Field Optimization (Koehl & Delarue, J Mol Biol. 1994; 239: 249-275) and Locally Enhanced Sampling (Roitberg & Elber, J Chem Phys 1991; 95: 9277-9287).

Combinatorial Searches (Dunbrack & Karplus, J. Mol. Biol. 1993; 230: 543-574; Tuffery et al., J. Biomol. Struct. Dynam. 1991; 8: 1267-1289; Wilson et al., J. Mol. Biol. 1993; 229: 996-1006) are employed on discrete conformers and may be followed by a continuous minimization in the final stage of refinement. It should be noted that there is no guarantee that any of the above will converge to a valid solution. Another widely used method is Dead End Elimination (DEE). It is based on the identification of rotamers that are absolutely incompatible with the global minimum energy conformation, eliminating rotamers that cannot contribute to local energy minima of a certain or higher order. Conformations comprising such rotamers can be qualified as dead ending (Desmet et al., Nature 1992; 356: 539-542; Desmet et al., FASEB J. 1997; 11: 164-172; Lasters & Desmet, Protein Eng. 1993; 6: 717-722). If enough rotamers can be eliminated by recursive application, the global minimum can be found (Goldstein, Biophys J. 1994; 66: 1335-1340). DEE can not, however,

find a population of low energy solutions.

The A* algorithm finds the optimal path from the root node to a goal node in a search tree using a cost function labeled f* (Leach & Lemon, Proteins 1998; 33: 227-239). Each node has a unique f* value composed from the cost of searching the node from the start node, and the estimated cost of reaching the goal node. f* is optimized in an iterative manner: the node with the smallest value of f* is expanded and new values of f* are calculated for its successor node.

The optimal method known so far for identification of proteins' low energy side chain conformations is a combination of DEE with the A* algorithm, which has been employed for constructing partition functions. The A* algorithm approach may find the best N solutions, but it is restricted to relatively small proteins. The largest protein solved by this algorithm so far contained 68 amino acids, which comprise about 10^{43} combinations - depending on the complexity of the rotamer library - while proteins with a much larger number of combinations are common. As a "stand alone" algorithm (without the DEE preprocessing stage) the A* algorithm reaches a maximum of 10^{21} combinations. For an effective search by the A* algorithm, it must have a good estimate of the cost to reach a goal node. This is problematic due to interactions between residues that have not yet been assigned. Those limitations raise the need for a novel robust algorithm that finds the global minimum and the lowest energy conformations in larger systems. Unfortunately, such an algorithm is not currently available.

Third Biological Problem: Prediction of Loop Structure

As previously noted, prediction of the structure of proteins requires a search in combinatorial space, which currently has no suitable solution. The prediction of protein structure can itself be divided into a number of smaller problems, which in themselves also require searches in combinatorial space. One example of such a problem is the very complicated prediction of loop structure.

Structural genomics projects are employed to provide an experimental structure or a good model for newly discovered sequences that emerge from the various genome projects. Brenner & Levitt (Protein Sci 2000; 9: 197-200) suggest, based on an analysis of sequence similarity databases, that the number of new folds is diminishing gradually, so that most common folds may soon be known. Homology modeling may thus become a prevailing tool for predicting a 3-dimensional structure of a protein sequence, if it shows a reasonably high

sequence similarity to another protein (a "template") with a known tertiary structure. In that case, secondary structure elements are transferred from the template to the target protein. However, stretches of "loops" or "coils" remain undetermined and must be predicted. Homology modeling has been employed for quite a while with some success. Lessons from milestone predictions such as that of HIV-1 protease based on the aspartic protease of Rous sarcoma virus (Weber, Science 1989; 243: 928-931) and amyloid precursor protease inhibitor domain from bovine pancreatic trypsin inhibitor (Struthers et al., Proteins 1991; 9: 1-11) have been useful for subsequent homology modeling for a plethora of structures. Most of the predicted loops are variable in both length and sequence in a family of proteins, and thus require a process of 3-dimensional insertions and deletions for their modeling. Even if such sequence and length mismatches are localized to short segments of the protein, there may still be significant rearrangements of the backbone conformations. Other regions of the two proteins, which are highly similar in both sequence and length, are termed "framework" regions. The non periodic structures that connect two sequential secondary structures are termed "loops" and described as having an "irregular conformation" or "random coil" (Oliva et al., J. Mol. Biol. 1997; 266: 814-830). The construction of loops - finding appropriate coordinates for variable regions in homology construction of proteins with insertions and deletions - is an important problem in globular proteins (Alwyn & Thirup, EMBO J. 1986; 5: 819-822 ; Brooks et al., J. Comput. Chem. 1983; 4: 187-217; Bruccoleri et al., Nature 1988; 335: 564-568; Bruccoleri & Karplus, Biopolymers 1987; 26: 137-168; Geer, Proteins 1990; 7: 317-334; Palmer & Scheraga, J. Comp. Chem. 1991; 12: 505-526; Summers & Karplus, J, Mol. Biol. 1990; 216:991-1016). The construction of loops is also a significant problem in other fields of structural biology. It is an immensely complex combinatorial problem: one should find fragments that should be properly inserted between the two end points of a loop, and subsequently evaluate their energies.

Extensive structural studies were employed on immunoglobulins, which are nearly identical in sequence and structure but differ dramatically in their binding specificities (Bruccoleri et al., Nature 1988; 335: 564-568; Chothia et al., Science 1986; 233: 755-758; Chothia & Lesk, J. Mol. Biol. 1987; 196: 901-917; Fine et al., Proteins 1986; 1: 342-362; Tramentano & Lesk, Proteins 1992; 13: 231-245). The specificity is due to six hypervariable loops, called "complimentarity determining regions" (CDR's). Understanding the structure of these loops may teach us a great deal about antigen binding, catalytic antibodies, and molecular recognition.

Another example are the intracellular loops that connect transmembrane helices in G-Protein Coupled Receptors (GPCRs) and form a potential ligand binding domain on the extracellular side, or the G-protein binding domain on the intracellular side of the membrane (Kazmi et al., Biochemistry 2000; 39: 3734-3744; Heymann et al., J. Struct. Biol., 1999; 128: 243-249). The prediction of their 3-dimensional conformation is crucial for understanding the mechanism (Gather et al., Nature 1993; 362: 345-348; Kyle et al., J. Med. Chem. 1994; 37: 1347-1352; Cypess et al., J. Biol. Chem. 1999; 274: 19455-19464; Zhang et al., Protein Sci. 1999; 3: 493-506; Lee et al., J. Biol. Chem. 2000; 275: 9284-9289) and for the subsequent effort to design GPCR related drugs (Mukherjee et al., J Biol. Chem. 1999; 274: 12984-12989).

Modeling of chemical rings raises the same problem where the constraints emerge from the need to close the ring with chemically reasonable bond lengths and angles (Go & Scheraga, Macromolecules 1970; 3: 178-187; Bruccoleri & Karplus, Macromolecules 1985; 18: 2767; Shenkin et al., Biopolymers 1987; 26: 2053-2085; Palmer & Scheraga, J. Comp. Chem. 1991; 12: 505-526).

Many current strategies divide the problem into two sub problems. First, one has to find geometrically acceptable conformations for polypeptide backbone fragments of correct length to be inserted between the two end points of a loop, within a framework of a known protein structure (Go & Scheraga, Macromolecules 1970; 3: 178-187; Weiner et al., J. Amer. Chem. Soc. 1984; 106: 765-784 ; Bruccoleri & Karplus, Macromolecules 1985; 18: 2767; Shenkin et al., Biopolymers 1987; 26: 2053-2085). This step usually yields multiple solutions. Second, the correct polypeptide among the suggested solutions in the first stage is selected usually by energy criteria.

Methods relying on a grid search for most of the backbone dihedral angles of the loop (Bruccoleri & Karplus, Biopolymers 1987; 26: 137-168; Moult & James, Proteins 1986; 1: 146-163), where the initial grid points are chosen from allowed regions of the Ramachandran map were suggested. These allowed regions were determined from the distribution of backbone torsion angles observed in a set of protein structures that have been determined by high resolution crystallography. A grid search grows exponentially in the number of degrees of freedom of the system, and therefore, such a method is restricted to relatively short loops.

Database search methods were initially proposed by Jones & Thirup (EMBO J. 1986; 5: 819-822) and extended by Summers & Karplus (J. Mol. Biol. 1990; 216: 991-1016). A database of high-resolution X-ray structures is scanned for amino acid segments with similar

geometric descriptors and size as the desired loop. The selected segments are docked into the protein, and geometric and energy criteria are used to determine their viability. Koehl & Delarue (Nat. Struct. Biol. 1995; 2: 163-170) employed a database search scheme combined with a self-consistent mean field approach to add the side chains. The loop length was shown to pose a limit for a reliable database search: as Summers & Karplus (J. Mol. Biol. 1990; 216: 991-1016) pointed out, this method is limited for up to six residues in length. In addition, a relatively large database of well-solved structures is required because it needs to cover most of the test cases. Deane & Blundell (Proteins 2000; 40: 135- 144) have recently employed an exhaustive ab initio search over computer-generated fragments to generate up to 8 residues loops.

In tree search methods (Brooks et al., J. Comput. Chem. 1983; 4: 187-217; Bruccoleri et al., Nature 1988; 335: 564-568) the search strategy is based on nodes, which may be expanded during the conformational search. The yield rate is low for structures of moderate size, and prohibitively low for large structures (Shenkin et al., Biopolymers 1987; 26: 2053-2085).

In the random tweak method (Shenkin et al., Biopolymers 1987; 26: 2053-2085), unconstrained structures are generated, in which all the dihedral angles are set to random values. Loop constraints are subsequently enforced geometrically by collectively altering ("tweaking") all the dihedral angles in an iterative process. Fine et al. (Proteins 1986; 1: 342-362) also generated a large number of random conformations for the backbone of the desired loop, followed by minimization and/or molecular dynamics with the remainder of the molecule held fixed. Kyle et al. (J. Med, Chem. 1994; 37: 1347-1352) employed a combination of homology modeling of the known transmembrane structure of bacteriorhodopsin, with energy minimization, molecular dynamics, and a two stage conformational search for a docking simulation. These techniques search conformational space in the vicinity of the starting point for a local energy minimum or minima.

The bond scaling-relaxation procedure meets the geometric and energy requirements simultaneously (Zheng et al., J. Comp. Chem. 1993; 14: 556-565). Random initial conformations are generated with standard bond lengths and angles. Bond lengths for each initial conformation are scaled to meet the loop-constrained distance, and systems are relaxed to a local energy minimum. This method was later enhanced by combining a multiple copy sampling method (Zheng et al., Protein Sci. 1993; 2: 1242-1248 ; Zheng et al., Protein Sci. 1994; 3: 493-506). The improved method was employed to handle loops with up to 12

residues.

The critical point in all these methods is in the difficulty to provide many different closures which cover a large conformational space that is required in order to maximize the probability that the correct (or very close) structure may be included (Zheng et al., J. Comp. Chem. 1993; 14: 556-565). Current procedures are either limited to small loops or, they explore only a fraction of the conformational space. Thus, potentially good solutions may be overlooked. There is a need for more efficient search strategies that explore the entire conformational space to find all possible conformations that obey loop closure geometric criteria. Those solutions can be subsequently evaluated by more refined criteria.

Other Biological Problems

In addition, there are many other biological problems which may be considered as requiring a combinatorial search, such as those associated with rational drug design.

A fundamental assumption for rational drug design is that drug activity is obtained through the molecular binding of one molecule (the ligand) to the pocket of another, usually larger, molecule (the receptor, commonly a protein). In their active, or binding, conformations, the molecules exhibit geometric and chemical complementarity, both of which are essential for successful drug activity. By binding to these macromolecules, drugs may modulate signal pathways, for example by altering sensitivity to hormonal action, or by altering metabolism, for example by interfering with the catalytic activity of the enzyme. Most commonly, this is achieved by binding in the specific cavity of the enzyme (the active site) which catalyses the reaction, thus preventing access of the natural substrate(s). In other cases, such as the transmembrane proteins, an "antagonist" may be designed in order to prevent the binding of an "agonist" (the natural molecule that activates the signal transduction) or, in case of reduced biological response, a stronger binding agonist may be required as a drug.

The modeling of molecular structure is a complex task, in particular because most molecules are flexible, being able to adopt a number of different conformations that are of similar or close energy. The modeling of the binding process is also a difficult task, as the characteristics of the receptor, the ligand, and the solvent in which these are found have to be taken into account. Although chemists strive to obtain models that are as accurate as possible, several approximations have to be made in practice. It is clear that the more accurate the model used, the better the chances chemists stand in predicting molecular

interactions. Nevertheless, a large number of predictions made with approximate models have been confirmed with experimental observations. Recently, a few drugs have been designed by computer theoretical methods. This has encouraged researchers to build tools that use approximate models and investigate the extent to which these tools can be useful. These approximate models pose difficult algorithmic questions. More accurate molecular modeling, gained through better theoretical understanding or increased computational power, can only improve the techniques developed with simpler models.

Depending on whether the chemical and geometric structure of the receptor is known or not, the problems that arise can be classified into two broad categories. If the receptor is known, chemists are interested in finding if a ligand can be placed inside the binding pocket of the receptor in a conformation that results in a low energy for the complex. This problem is referred to as the docking problem. It has several variations: an accurate description of the binding interaction may be desired, or an approximate estimate may be sought of which ligands, from those contained in a huge database, are likely to fit inside the receptor.

Very often the binding pocket is unknown. In fact, the 3D structure of relatively few large molecules (or macromolecules) has been determined by X-ray crystallography or NMR techniques, although this number is increasing rapidly. In this case, indirect approaches must be adopted, which use a number of ligands that interact with that specific receptor. These ligands have been discovered mainly by experiments. Using the geometric structure and the chemical characteristics of these molecules, chemists attempt to infer information about the receptor. In particular, chemists are interested in identifying the pharmacophore present in these ligands. The pharmacophore is a set of features in a specific 3D arrangement contained in all the active conformations of the considered molecules. A prevailing hypothesis is that the pharmacophore is the part, or parts, of the molecule that is responsible for drug activity, while the rest of the molecule is a scaffold for the pharmacophore's features. If the pharmacophore is determined, by examining the different activities, relative shapes, and chemical structures of the initial molecules, chemists can use it to design a more potent pharmaceutical drug.

The techniques that have been used so far in computer-aided drug design include robotics (kinematics and planning), graphics algorithms (visualization of molecules), geometric calculations (surface computation), numerical methods (energy minimization), graph theoretic methods (invariant identification), randomized algorithms (conformational search), computer vision methods (docking), and a variety of other techniques like genetic

algorithms and simulated annealing. A number of tools for performing complex geometric and energy calculations are now available and the success of these computer-aided methods is under evaluation.

The other part of the general problem of drug design emerges from the biomolecular targets. Advances in genomics, proteomics and bioinformatics are yielding new therapeutic targets for drug discovery efforts at a rapid rate. Given the virtually limitless numbers of compounds which could be tested for activity against these targets, biopharmaceutical research is becoming increasingly reliant on a synergistic approach for accelerating drug discovery. Novel computational methods are required in order to improve our ability to deal with the plethora of information that emerges from sequences of new proteins, in order to transform sequences into structures. Further, even detailed structural studies of proteins are limited in information content because they produce, at best, a limited set of static conformations, while in most X-ray studies, single structures emerge, that lack important information about proton positions of the protein and the solvent. Even once the full structures are known, the design of candidate drugs that may interact with these targets is an extremely difficult task. The field of structure based drug design is thus in great need of improved methods that would ease the above tasks.

The toughest limitation in most of these problems is the high level of their complexity, due to the large number of variables. Any search for "real" molecular structures requires a process of "optimization" of this large number of variables. Such a complex potential energy surface (PES) has many minima, of which one is the "global minimum" and is probably related to the native structure of the molecule.

Despite considerable recent progress, the general problem of global optimization remains unsolved (Wales & Scheraga, Science 1999; 285: 1368). Conventional minimization techniques are time consuming and tend to converge to local minima. Sampling of the "phase space" of biomolecules (Berne & Straub, Curr. Opin. Struct. Biol. 1997; 7: 181) may be helpful for searching regions of minima and for reducing the time required to reach such regions. Some of the main global minimization algorithms are presented here by order of decreasing applicability towards computer aided biological applications.

Protein structure prediction can be shown to be an NP-hard problem; the number of conformations grows exponentially with the number of residues. The native conformations of

proteins occupy a very small subset of these, hence an exploratory, robust search algorithm is required.

Simulated Annealing (SA) is a generalization of a Monte Carlo method that has been used for examining the equations of state and frozen states of n-body systems (Metropolis, J. Chem. Phys. 1953; 21: 1087). In an annealing process a melt, initially at high temperature and disordered, is slowly cooled. As cooling proceeds, the system becomes more ordered and approaches a "frozen" ground state at T=0. An initial configuration is perturbed and the change in energy dE is computed. If the change in energy is negative the new configuration is accepted. If the change in energy is positive it is accepted on the basis of the Boltzmann factor exp-(dE/kT). This process is then repeated sufficient times to give good sampling statistics for the current temperature, and then the temperature is decreased and the entire process repeated until a frozen state is achieved at T=0. SA is suitable for optimization problems of large scale (Holm & Sander, Proteins 1992;14: 213; Lee & Subbiah, J. Mol. Biol. 1991; 217: 373; Hwang & Liao, Protein Eng. 1995; 8: 363; Press et al., Numerical Recipes, Cambridge University Press, New York, NY, 1986; 326), especially ones where a desired global minimum is hidden among many, much poorer, local minima.

Genetic Algorithms (GAs) have been applied to a number of optimization problems with some success (Tuffery et al., J. Comput. Chem. 1993; 14: 790). GAs take their inspiration from the Darwinian principle of evolution: natural selection and survival of the fittest (Forrest S (1993); Science 261, 872). Each iteration of GAs involves a competitive selection that weeds out poor solutions. The solutions with high "fitness" are "recombined" with other solutions by swapping parts of a solution with another. Solutions are also "mutated" by making a small change to a single element of the solution. GAs are simple, tend not to get "stuck" in local minima and can often find a globally optimal solution. No derivatives or any other problem-specific calculations need to be done. However, there is no guarantee that it will converge to a valid solution, and many iterations are needed in order to achieve convergence criteria.

Taboo Search (TBS) (Glover, Computers and Operations Research 1986; 5: 533) is superior to SA both in the time required to obtain a solution and the quality of the latter (Cvijovic & Klinowski, Science 1995; 267: 664). At initialization the goal is to make a

rough examination of the solution space, but as candidate locations are identified the search is more focused to produce local optimal solutions. TBS is problem independent and can be applied to a wide range of tasks. It is very easy to implement and the entire procedure occupies only a few lines of code. It is conceptually much simpler than SA and GA. However, it cannot guarantee to solve the multiple minima problem in a finite number of steps, and may require long computing times.

The group of H. Scheraga has been very active in devising methods for global optimization. Potential function deformation and smoothing methods (Piela et al., J. Phys. Chem. 1989; 93: 3339: Pillardy & Piela, J. Phys. Chem. 1993; 99: 11805; Pillardy et al., J. Phys. Chem. 1999; 103: 7353) transform the energy "landscape" of a biomolecule and enable a study of those parts of the PES that are more relevant for finding the global minimum. However, the deformed surface may include too many "catchment basins" while smoothing by the diffusion equation method does not guarantee the isolation of the lowest energy minimum in multi-dimensional problems. Conformational space annealing (Lee et al., J. Comput. Chem. 1997; 18: 1222), which narrows the search on a full conformational space to regions of low energies and starts a search with a "pool" of minimized conformations, that are later modified by picking random variations from the "pool", is also limited to a small number of variables.

Dead End Elimination (DEE) is based on identifying solutions that are absolutely incompatible with the global minimum. (Desmet et al., Nature 1992; 356: 539; Lasters et al., J. Prot. Chem. 1997; 16: 449). Solutions that cannot contribute to local energy minima of a certain or higher order are eliminated. One should write an energy (cost) function as a sum of terms which are themselves functions of maximally two variables. A value for the i-th variable x_i cannot be consistent with the globally optimal solution if another value for the same variable, x'_i, can be found so that:

$$\mathbf{(1)}\ c(x_i) + \Sigma_{j=1,N}\ \min c(x_i, x_j) > c(x'_i) + \Sigma_{j=1,N}\ \max c(x'_i, x_j)$$

When the process iterates, enough solutions are eliminated and the global minimum can be found (Goldstein, Biophys. J. 1994; 66: 1335). Other methods combine a discrete search strategy with a continuous minimization in the final stage of refinement (Dunbrack & Karplus, Mol. Biol. 1993; 230: 543; Vasquez, Biopolymers 1995; 36: 53). The most popular application for this algorithm is the determination of side chain conformation of

proteins. If the DEE-method fails to reach a unique structure, an additional step is required such as a brute force combinatorial search or a clustering approach on the remaining conformations (Becker, Proteins 1997; 27: 213). DEE faces a serious practical problem: It can minimize on a one per one basis, while the above condition requires minimization over all possible values. An additional disadvantage is the inability to find the ensemble of low energy solutions.

Statistical Methods (SMs) employ a model of the objective function to bias the selection of new sample points. These methods are justified with Bayesian arguments that suppose that the particular objective function to be optimized comes from a class of functions that are modeled by a particular stochastic function (Mockus, J. Global Optim. 1994; 4: 347). Information from previous samples of the objective function can be used to estimate parameters, and this refined model can subsequently be used to bias the selection of points in the search domain. The problem in using statistical SMs is whether the statistical model is appropriate for a problem. Additionally, it is difficult to write computer codes for high dimensional optimization problems due to the mathematical complexity. Many times, SMs rely on dividing the search region into partitions, which limits these methods to problems with a moderate number of dimensions.

Unfortunately, none of the above attempted solutions is able to provide a suitable answer to the above specific problems within the larger problem of protein structure prediction, let alone to provide a more general solution for searches within combinatorial space.

There is therefore a need for, and it would be useful to have, a solution for searches in combinatorial space, which would be efficient, rapid and simple in execution, and which would be useful for these different types of biological problems, such as problems within the larger problem of the prediction of protein structure.

SUMMARY OF THE INVENTION

The present invention discloses a system and method for searching through combinatorial space, without a combinatorial explosion. At a more basic level, each combination may be considered to consist of *variables*, each of which may assume at least one *value*. According to the present invention, each variable preferably assumes one value of

a set of discrete values, although alternatively, each variable may assume one value out of a range of continuous values or out of a function, for example. These variables interact with each other in a manner which is known for each individual interaction. Preferably, individual interactions can be described for pairs of variables, such that the interactions are pairwise interactions. The search is performed by sampling one *value* of each *variable* to obtain a combination. This process is then repeated, typically many times. Each combination is evaluated by a quantitative measurement. The quantitative measurement is preferably a cost function, for which the desired outcome is generally maximized or at least increased during the process of determining which combinations best fulfill the cost function. For example, if the cost function is an energy minimization function, then the combinations are preferably selected which have lower energy costs or values.

The present invention then attempts to determine which elements do not contribute to combinations which provide at least some minimum desired value for the quantitative measurement, and/or which contribute to combinations which provide a value for the quantitative measurement which is below some cut-off or threshold for desirable values. In other words, these elements do not contribute toward the "best" or most satisfactory combinations for the system. These elements are then preferably eliminated or at least segregated from the remaining possible elements for forming the combinations.

The process of evicting values of *variables* is preferably repeated until a predetermined number of combinations remain, which consist of the elements which have not been eliminated and/or segregated. At this point, an exhaustive search is most preferably performed, according to the quantitative measurement and/or according to some other measurement parameter or parameters.

The present invention accomplishes these tasks by examining each value of a basic element at least once, and preferably a plurality of times, during the search process. Therefore, each value can be said to be searched in an exhaustive search, yet not every combination of the values for the basic elements needs to be exhaustively searched. Thus, the present invention combines the efficiency of non-exhaustive, stochastic search processes with the efficacy of exhaustive searches.

According to the present invention, there is provided a method for searching through combinatorial space, the space featuring multiple combinations, each combination being composed of at least one element, the steps of the method being performed by a data processor, the method comprising the steps of: (a) providing a quantitative parameter for

determining success of a result of a search through the combinatorial space, said quantitative parameter being measurable for each combination; (b) dividing the combinations in the combinatorial space into ensembles, each ensemble featuring at least one combination; (c) calculating a value for said quantitative parameter for at least one combination of each ensemble; (d) determining an effect of each element on said value of said quantitative parameter; and (e) retaining at least one combination according to said effect, to provide a result of searching through the combinatorial space.

Hereinafter, the term ""amino acid" refers to both natural and synthetic molecules which are capable of forming a peptide bond with another such molecule.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is herein described, by way of example only, with reference to the accompanying drawings, wherein:

FIG. 1 is a flowchart of an exemplary method according to the present invention;

FIG. 2 is a schematic block diagram of an exemplary system according to the present invention;

FIG. 3 shows a flow chart for the hydrogen positioning algorithm;

FIG. 4 shows a molecule that contains two carbonyls, one sp^2 amide and one hydroxyl, that form together a single ensemble. The two carbonyls (1,2) act as acceptors, the hydroxyl donates one non trivial hydrogen (3) and two non trivial lone pairs (4,5), and the amide donates one trivial hydrogen (6). Atom 3 and lone pairs 4, 5 are one segment because they are bonded to the same oxygen;

FIG. 5A shows an exemplary initial 2D matrix for the system in Figure 4. The hydroxyl hydrogen (3) can form a hydrogen bond to any of the carbonyls (1,2) and the hydroxyl lone pairs (4, 5) can form a hydrogen bond to the trivial hydrogen (6). FIG. 5B shows the refined 2D matrix. The hydroxyl two lone pairs are degenerate, therefore one of them can be omitted (5->6). The omitted lone pair is automatically added after the hydrogen and first lone pair are located. FIG. 5C: using the 2D matrix, a 3D matrix is formed to keep all the possible combinations. Each combination is evaluated, and the best combination is the result;

FIG. 6 shows an example of a "big" system. The initial 2D matrix in case of a large biological system (for example, a protein). An attempt to create the 3D matrix will exceed the computer capabilities. Therefore, the 2D matrix is refined by evicting high energy

components;

FIG. 7 shows a "test" protein with 1186 amino acids: 13 are serines (those are marked as CPK model) (13 segments) and 1173 glycines (0 segments). The stochastic search began with a total number of $5.02*10^{10}$ combinations and reached $2.7*10^{3}$ combinations after 204 iterations, which were then evaluated exhaustively. The global minima for hydrogens' positions was found;

FIG. 8 shows a graph of the natural log of (total number of possible combinations) vs. the iteration number in the pure "stochastic approach". Five proteins are presented;

FIG. 9 shows a graph of energy distribution in the 1^{st} and 4^{th} iterations for 5PTI (A), 5RSA (B), 2MB5(C), 1NTP(D);

FIG. 10 shows a Ribbon display of trypsin (1NTP) and its polar residues. Many polar hydrogens create hydrogen bonds with water molecules. However, no water molecules' coordinates are included in the PDB file;

FIG. 11 shows a model of crambin (46 amino acid residues) as a test case for comparison of a full exhaustive search to a stochastic search in finding the 10,000 lowest energy conformations. The backbone of crambin is presented as a ribbon. The non hydrogen atoms are presented by ball and stick models;

FIG. 12 shows a comparison of stochastic and exhaustive searches in finding lowest energy conformations for 1-10,000 conformers. The % deviation between the two searches is on the lowest curve;

FIG. 13A shows a percentage of angles in E. coli ribonuclease HI that may be detected: Out of 115 dihedral angles, 7 angles are missing from the rotamer library; Figure 13B shows a percentage of angles in E. coli ribonuclease HI that were detected by the stochastic algorithm;

FIG. 14 shows values of α for 2 to 29 possible rotamers of a single residue that lead to elimination with high probability. Each number of rotamers has an associated value of α(triangles). The larger the number of rotamers, the smaller is α. For each given number of rotamers and α, the % certainty is calculated (squares);

FIG. 15 shows an example of a 6 residues (0-5) loop. Residues 0 and 5 are part of the transmembrane helix. A search is performed for the conformation of residues 1-4. The method of the present invention is employed to explore the conformational space of the loop to find all possible loop closure conformations defined by equation 2;

FIG. 16 shows the dihedral angles definition: ψ of a residue n, in the construction

strategy, is the ψ of the previous residue toward the N-terminal;

FIG. 17 shows the 10,000 "lowest cost function" conformations in a 4 residues' test case. A stochastic and an exhaustive search achieved the same global minimum. The 66 first conformations are identical.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention discloses a system and method for searching through combinatorial space, without a combinatorial explosion. The search is performed for various combinations of basic elements, according to at least one desired property of the combination, which is translatable into a quantitative measurement of the success of the search. The present invention then attempts to determine which elements do not contribute to combinations which provide at least some minimum desired value for the quantitative measurement, and/or which contribute to combinations which provide a value for the quantitative measurement which is below some cut-off or threshold for desirable values. Preferably, those elements are selected which only contribute to combinations which fail to meet the minimum threshold for desirable values. In other words, these elements do not contribute toward the "best" or most satisfactory combinations for the system. These elements are then preferably eliminated or at least segregated from the remaining possible elements for forming the combinations.

The process of sorting through the elements is preferably repeated until a predetermined number of combinations remain, which consist of the elements which have not been eliminated and/or segregated. Such a predetermined number is optionally and more preferably an actual numerical value for the total number of combinations, but alternatively may be a threshold for the minimum desired value for the quantitative measurement which the combination must satisfy to be included in the remaining combinations. At this point, an exhaustive search is most preferably performed, according to the quantitative measurement and/or according to some other measurement parameter or parameters.

Since there may be a huge number of combinations, preferably samples of combinations are examined with regard to the effect of the elements on the quantitative measurement for evaluating the combination. Those elements of the combinations which have consistent maximization and/or promotion of the quantitative measurement are then kept, while the other elements are dropped. This process is then repeated until some maximum number of combinations is found, which could then optionally be further

evaluated according to a similar parameter and/or some other parameter or characteristic. Such a group of combinations can also optionally be viewed as a population of combinations having a particular minimum value for the quantitative measurement.

At a more basic level, each combination may be considered to consist of *variables*, each of which may assume at least one *value*. According to the present invention, each variable preferably assumes one value of a set of discrete values, although alternatively, each variable may assume one of a range of continuous values or out of a function, for example. These variables interact with each other in a manner which is known for each individual interaction. Preferably, individual interactions can be described for pairs of variables, such that the interactions are pairwise interactions. The quantitative measurement of the combinations of variables is preferably a cost function, for which the desired outcome is generally maximized or at least increased during the process of determining which combinations best fulfill the cost function. For example, if the cost function is an energy minimization function, then the combinations are preferably selected which have lower energy costs or values.

Different properties can optionally be used in order to create the cost function for performing the search in combinatorial space. For example, for searching for different protein structures according to an amino acid sequence, the cost function could optionally be the energy minimization of the combination, such that the selected structure would represent an energy minimum or near-minimum. Such an energy cost function is also useful for more specific or "sub" problems within the larger problem of protein structure prediction. For example, minimization of the predicted location of polar protons and side chains for amino acids also provides a useful quantitative parameter for these types of combinatorial searches. It should be noted that in this case, maximization of the desired quantitative parameter is actually achieved through *minimization* of the value of the energy calculation for the combination.

However, it should be noted that substantially any cost function could optionally be used with the method of the present invention. The cost function would not even necessarily need to be related to a biological problem, but could instead be related to other types of problems, such as optimization of a cost function for monetary value (for a literal, financial "cost"), for example.

The present invention accomplishes these tasks by examining each value of a basic element at least once, and preferably a plurality of times, during the search process.

Therefore, each value can be said to be searched in an exhaustive search, yet not every combination of the values for the basic elements needs to be exhaustively searched. Thus, the present invention combines the efficiency of non-exhaustive, stochastic search processes with the efficacy of exhaustive searches.

As previously noted, an additional exhaustive search may optionally be performed after the execution of the present invention, for example in order to identify the absolute minimum as well as a plurality of local minima. Such an additional exhaustive search is particularly preferred when the initial search process according to the present invention includes a stochastic search and/or comparison component, which is the preferred embodiment of the present invention. It should be noted that the present invention is clearly distinguished from background art search methods in a number of respects. First, the present invention is not based upon, nor is it a modification of, any of the known methods in the art. Second, each value of every variable in the combinatorial search space must be probed to determine whether it should be evicted from the search space, unlike other stochastic search methods, which can not guarantee the probing of each and every value in the combinatorial search space. Third, the present invention is also optionally and preferably able to obtain a population of local minima in addition to the global minimum. Yet, as described in greater detail below, the present invention is able to accomplish these goals with a stochastic search, while providing the efficacy of the exhaustive search, as proven below by a comparison of the results of the present invention with the results of full exhaustive searches used alone.

The principles and operation of the present invention may be better understood with reference to the drawings and the accompanying description, which are provided through several sections. The first part of the description (in this section) centers around an exemplary general method according to the present invention, and a basic exemplary system for implementation thereof. The subsequent sections refer to specific biological problems, and are labeled with the name of each type of problem. These sections are intended to describe examples for suitable implementations and applications of the present invention, and are not otherwise intended to be limiting in any way.

Referring now to the drawings, Figure 1 is a flowchart of an exemplary but preferred general method according to the present invention for searching through combinatorial space. As shown, in step 1, the combinatorial space is provided. Such a combinatorial space features multiple combinations of basic elements. The combinatorial space is optionally created, for example by creating multiple structures having the basic elements according to

some pattern, plan and/or scheme. Alternatively, the combinatorial space may optionally have been previously defined. For example, for biological problems, the combinatorial space may already be defined according to the type of biological structure which is to be analyzed.

In step 2, according to preferred embodiments of the present invention, each combination is optionally and preferably constructed from *variables*, each of which may assume at least one *value*. According to the present invention, each variable more preferably assumes one value of a set of discrete values, although alternatively, each variable may optionally assume one out of a range of continuous values or out of a function, for example. These variables interact with each other in a manner which is known for each individual interaction. Preferably, individual interactions can be described for pairs of variables, such that the interactions are pairwise interactions.

In step 3, the quantitative parameter is determined, according to which the success of the search is measured. The quantitative parameter must be measurable for each combination of the combinatorial space. Typically, the quantitative parameter is calculated according to the basic elements of each combination, optionally with the additional consideration of the effect of structural features and/or interactions on this measurement. For biological problems, such as protein structure prediction, the type of quantitative parameter for examining the particular problem may already be known. For example, for the prediction of the locations of polar protons within a protein, the best quantitative parameter is preferably the energy minimization for the combination, determined according to equations which are known in the art and which are described in greater detail below with regard to Section 1.

The quantitative measurement of the combinations of variables is preferably a cost function, for which the desired outcome is generally maximized or at least increased during the process of determining which combinations best fulfill the cost function. For example, if the cost function is an energy minimization function, then the combinations are preferably selected which have lower energy costs or values.

In step 4, the contribution of each element or variable is evaluated, to determine the effect of particular elements or values of particular variables of each combination on the quantitative parameter or cost function. Such an effect is preferably determined through both the values of the variables, and the interaction between these variables, as assessed through the cost function.

The preferred effect is for consistent maximization of the cost function. Consistent maximization is optionally measured according to the distribution of values of the cost

function for a large group of combinations or "configurations" of the whole set of variables. According to preferred embodiments of the present invention, particularly if large numbers of variables are involved, preferably the effect of these different values is determined through a stochastic analysis, since an exhaustive analysis could prove to be prohibitively inefficient and time-consuming. The stochastic analysis is preferably performed by randomly selecting values for each variable in order to form a combination, more preferably in order to form a plurality of different combinations. Most preferably, a predetermined number of such combinations are formed as part of a sampling process. The outcome or value of the cost function for each combination is then calculated, according to both the values of the variables and the interaction between these variables.

In step 5, optionally and preferably, those elements or values of variables are removed which do not contribute to consistent maximization of the desired outcome of the cost function, as previously described. More preferably, those values of variables are removed which contribute only to less desirable outcomes of the cost function, or outcomes which fall below a certain minimum threshold, and not to any outcomes which are above a certain threshold for desirable outcomes. For example, for a cost function involving energy minimization, those values for variables are preferably removed which are found only in combinations for which the energy cost is higher than a certain threshold (less desirable outcome), but not in combinations for which the energy cost is below another, low energy threshold (more desirable outcome). Alternatively, rather than being removed, these values can optionally be "marked " and/or segregated, for example for further analysis.

In step 6, if the total number of combinations has reached some minimum value, then these combinations are optionally and more preferably further analyzed according to the cost function, and/or some other parameter to determine the results of the combinatorial search. Optionally, an exhaustive search could even be performed within the minimum number of combinations for the combination(s) of interest, again as evaluated according to the cost function, and/or some other parameter of interest. Such a group of combinations can also optionally be viewed as a population of combinations having a particular minimum value for a desirable outcome of the quantitative measurement.

Otherwise, steps 4 and 5 are preferably repeated, until this minimum number of combinations is reached. It should be noted that the "minimum number" could optionally refer to an absolute number of combinations, and/or a minimum value for the cost function which such combinations must meet as a "cut-off" threshold.

Figure 2 shows an exemplary system according to the present invention, for implementation of the method of Figure 1. As shown, a system **10** features a computational device **12**. In this embodiment, computational device **12** operates a number of functional modules, which collectively enable the method of Figure 1 to be executed. These functional modules are optionally and preferably implemented as software modules, but alternatively may implemented as hardware, firmware or a combination thereof.

As shown, one such module is a combination storage module **14**, which holds the combinations currently under consideration in their respective ensembles. A quantitative parameter calculation module **16** then calculates the value of the quantitative parameter for at least one combination in each ensemble from combination storage module **14**. An evaluation module **18** creates a plurality of samples of combinations from the elements of the combinations, and evaluates the effect of each element on the value of the quantitative parameter for the combination, such that certain elements are preferably retained as consistently contributing toward maximized values for the quantitative parameter for the combination. These modules preferably interact until a certain minimum number of combinations are held in combination storage module **14**, which represent the results of the search in the combinatorial space.

The preceding description generally illustrated the method and system of the present invention. The next Sections describe specific model systems which are handled by the present invention as specific problems, for which the present invention is able to provide a solution. These Sections include descriptions of searching through combinatorial space to locate polar protons (Section 1); locating amino acid side chains in proteins (Section 2); prediction of loop structure in proteins (Section 3); and other miscellaneous biological problems which are solved by the present invention (Section 4).

Section 1: Location of Polar Protons

The present invention is useful for solving the problem of correctly locating the polar protons within a biological molecule, such as a protein molecule or DNA, for example. The location of such polar protons in turn determines the location of hydrogen bonding, either within the biological molecule itself, or alternatively between the biological molecule and another molecule. This specific implementation of the present invention thus solves an important scientific problem.

The specific implementation of the present invention which is described in this section under "Methods" was also tested against other methods known in the art, as described under "Results". It should be noted that these methods and results are presented for the purposes of illustration only, and are not intended to be limiting in any way. The interpretation for these results is then discussed under "Discussion".

Methods

For the purposes of testing, the method of the present invention has been implemented as a computer software program, written in C++. It operates as illustrated in the flow chart of Figure 3. As shown, in step 1, the program optionally reads the Protein Data Bank coordinate file format (a PDB file), or alternatively receives the input information from another source. It uses auxiliary ASCII files which serve as databases to parametrize the system atoms. Those files contain the connectivity of all atoms, their charges, A and B parameters for the Lennard-Jones function, and bond lengths between hydrogens and heavy atoms. The user may add, delete and modify residue types easily by editing these files. These values are read from the file, or alternatively are input from another source, in order to parametrize the atoms in step 2.

In step 3, the hydrogens and lone pairs, which are about to be added, are divided into three categories: (1) Trivial hydrogens-those hydrogens that may be located using coordinates and hybridization of heavy atoms, such as aliphatic and aromatic hydrogens. (2) Non trivial hydrogens-polar hydrogens, which have rotational degrees of freedom, such as serine, threonine and tyrosine hydroxyls. (3) Non trivial lone pairs, which are those with the same geometrical properties of non trivial hydrogens.

Trivial hydrogens are added first, in step 4. Their coordinates are calculated using the coordinates of the heavy atoms, the bond length and angles from the database as well as the standard dihedral angles.

In step 5, non trivial hydrogens and lone pairs are divided into ensembles, and their coordinates are not yet calculated. An ensemble is defined as a group of non trivial hydrogens or lone pairs which interact among themselves. The ensemble cutoff is user defined. The user can assign a large ensemble cutoff value, and force the system to run as one big ensemble. The ensemble cutoff is measured from the coordinates of the heavy atom bonded to the non trivial atom, because the non trivial atom has not been located yet. Ensembles are composed of "segments". Each segment includes a rotation around a bond

connecting two heavy atoms, one of which is bonded to a polar proton. Each segment may employ various positions in space to fulfill H-bonding conditions.

In addition to the ensemble cutoff, two other cutoff conditions are optionally and preferably employed: an energy cutoff, in the usual sense of its use in non bonding energy calculations: the default is no cutoff. Another cutoff is used for locating hydrogen bonding partners around a rotatable segment (*vide infra*)-this may be smaller or larger than the "ensemble cutoff", however it should be always >3 Å to allow the inclusion of all close partners for H-bonding, and to avoid the risk of missing solutions for a segment. Increasing this cutoff over 4.5Å creates many non realistic optional partners and extends the time for searching solutions. The ensemble cutoff is employed for creating a group of relevant heavy atoms (hydroxyl oxygen, water oxygen, NH_3^+, amine, etc...) that must solve its relations with respect to all its members. Thus, if the cutoff is 4Å , it may well be that the distance between each pair of atoms A and B, or A and C, is smaller than 4Å , but $R_{B,C}$ may be > 4Å, while all three atoms are part of the same ensemble.

Each ensemble is preferably treated separately. In order to calculate the coordinates of the non trivial hydrogens and lone pairs, a two dimensional matrix is formed in step 6. It is a list of all hydrogen bonds that may be formed between donors and acceptors. The larger the H-bonding cutoff, the more options for hydrogen bond connections will be formed, and the larger the 2D matrix of alternative interactions will be.

As an example, the ensemble displayed in Figure 4 contains only two carbonyls (1,2), one amide and one hydroxyl, that form together a single ensemble. The hydroxyl donates one non trivial hydrogen (3) and two non trivial lone pairs (4,5), and the amide donates one trivial hydrogen (6). A segment is defined as a group of non trivial hydrogens and lone pairs bonded to a single heavy atom. For example, atom 3 and lone pairs 4 and 5 are one segment because they are connected to the same oxygen. Suppose the hydroxyl hydrogen (3) can form a hydrogen bond with any of the carbonyls, and the hydroxyl lone pairs (4,5) can form a hydrogen bond with the N-H, the full 2D matrix will have the form illustrated in Figure 5A. For forming a hydrogen bond to the amide the two lone pairs are degenerate, therefore one of them can be omitted for forming the initial alternative combinations of the 2D matrix (4->6 or 5->6). The omitted lone pair is automatically added after the hydrogen and first lone pair are located. Therefore, the initial 2D matrix will have the form illustrated in Figure 5B. The module refines the 2D matrix: a location that yields a high energy value ("bump") is deleted. The energy threshold is user defined, and non bonding energy expressions are employed.

Using the refined 2D matrix, a 3D matrix is formed in step 7, where all combinations in an ensemble are uniquely defined, i.e. in any combination there is only a single option for any non trivial (rotatable) hydrogen and non trivial lone pair. In the example, the 3D matrix has the form illustrated in Figure 5C. Each pair of lines constitutes one contribution. Each combination is evaluated, and the best combination is the result for the ensemble. In case of more than one ensemble the process is repeated for each ensemble.

The energy criterion used to evaluate the quality of each combination is a pairwise "non bonding" energy function: $E(r_{i,j}) = \sum_{i<j}(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} + \frac{q_i q_j}{\varepsilon * r_{i,j}})$, where $A_{i,j}$ is the repulsion parameter for the two (i, j) atoms, $B_{i,j}$ is their attractive polarizability parameter, q_i is the partial charge, and $r_{i,j}$ is the distance between atoms. ε is the dielectric constant chosen to be 4 in the tests of the algorithm. The code is flexible and the force field can be easily modified to any desired.

Energy calculations extend over the "borders" of each ensemble. The cutoff distance for calculating $E(r_{i,j})$ is user defined, however avoiding cutoffs is recommended so that long range electrostatic interactions may be accounted for. The main problem with this ensemble approach is to calculate interactions between non trivial atoms in one ensemble and non trivial atoms in another ensemble. In this case, the coordinates of the non trivial hydrogens in a second ensemble, that have not been positioned yet, are assumed to coincide with the coordinates of the heavy atoms to which they are bonded. This is a "unified atom" approximation justified by the relatively long distance between the known position of one atom and the yet undetermined position of another. However, the user can avoid this approximation by forcing the program to handle the system as one huge ensemble-in this case, all non trivial hydrogens and lone pairs are added simultaneously, with exact positions.

It is obvious that in case of a large biological system constituting a single ensemble, a very large combinatorial problem results. In RNase-A(5RSA), for example, there are $1.76*10^{59}$ alternative combinations for all rotatable hydrogens. An attempt to create the 3D matrix from the 2D matrix will exceed the computer capabilities. To reduce the size of the problem, a unique stochastic approach was developed. The algorithm switches from exhaustive to stochastic calculation of an ensemble once the number of combinations exceeds a user-defined threshold. In the ensemble, the locations in d_0 segments are unknown. For each non-trivial hydrogen or lone pair there is usually more than one location, but only one would give the lowest energy. Non trivial hydrogens and non-trivial lone pairs affect

each other: if a lone pair is located, its location will dictate the location of a hydrogen bonded to the same heavy atom, and vice versa.

Let $X=(X_1, X_2...X_{d0})$ be a configuration of d_0 segments in one ensemble. For each configuration X, the energy E=E(X) is calculated according to the energy function described above. The objective is to find the configuration which minimizes E. Since it is impossible to evaluate all the alternative configurations due to the large number of combinations, those steps are followed as an example for performing step 8, the evaluation of combinations:

1. Sample at random n configurations out of the large population of combinations $X_1=(x_{11}, x_{12}, \dots, x_{1d0}), \dots, X_n=(x_{n1}, x_{n2}, \dots, x_{nd0})$, where x_{11} is the first randomly picked conformation for the first segment and x_{n1} is the n^{th} randomly picked conformation for that segment. Figure 6A illustrates the first three configurations and the n^{th} configuration sampled from the 2D matrix. Compute the corresponding energy values: $E_1 = \sum_{j=1}^{d_0} e(1)j$ for configuration X_1, $E_n = \sum_{j=1}^{d_0} e(n)j$ for configuration X_n.

2. Construct the distribution F_E^n (n is of the order of 10^3). F_E^n is an assembly of energies that corresponds to n sampled configurations for the full protein. Define cutoff points H and L in F_E^n. H contains all configurations satisfying $E_i \geq F_E^n(1-\alpha)$, where $F_E^n(\alpha)$ is the α-th percentile of F_E^n, while L contains all configurations satisfying $E_i \leq F_E^n(\alpha)$. The number of configurations in each of H and L is $n_0=n*\alpha$. When n=1000 configurations and α=1% for highest and lowest energy configurations, $n_0=\alpha*n=0.01*1000=10$ so L=10 and H=10. In other words, H stands for the 10 highest energy systems (Figure 6B), while L stands for the 10 systems with the lowest energy.

3. Construct the vector h for the positions in configurations corresponding to the energies in H. The vector h is the element-wise intersection of all the configurations in H, in the following manner: if all configurations in H share the same value, say 5->1, at component j, (corresponding to x_{nj} of configuration X_n) then h_j=5->1; otherwise, h_j=0 (no common position for segment j in all high energy configurations.) For instance, in Figure 6B all configurations of H share the same values 5->1 and 23->34, therefore those configurations will be part of the vector h=(5->1, 23->34, .., 0). This is the vector constructed for n*α high energy configurations for d_0 segments indicating that the value 5->1 in segment 1, as well as the value 23->34 in segment 2 appear in all high energy configurations (figure 6C). No common position was found for the last segment d_0 in the high energy region.

4. Construct the vector l for the positions in configurations corresponding to the energies in L. The vector l is the union of all the configurations in L as illustrated in Figure 6D. Unlike vector h, more than one configuration may appear for each segment in l.

5. Compare h and l. If both h_j and l_j have a similar vector component, j, it will remain as a viable configuration for that segment, because it contributes also to low energy values. However, if $h_j \neq l_j$, than the corresponding segment component h_j will be evicted from subsequent iterations. It should be noted that in segments with size that equals 1 h_j will not be evicted from subsequent iterations because it is the only available solution. Figure 6E demonstrates the eviction of the value 5->1 from further calculations because it exists exclusively in the high energy vector, h, while the value 23->34 in segment 2 will not be evicted, because it also exists in vector l. The new 2D matrix will not contain the pair 5->1, as illustrated in Figure 6F. In order to avoid configurations with very high energy which might skew the results, the number of configurations n and the percentile value of α were chosen according to statistical formulae that deal specifically with the probability of justified and unjustified eviction of configurations from a large set of combinations. A minimization of incorrectly ruled out cases may be achieved by increasing α and n. However, the expected number of correctly ruled out cases also decreases, though, with a smaller slope. Values of n=500 and α=0.008 were chosen as a reasonable compromise. (step 8 of Figure 3)

6. Repeat steps 1 to 4 for the reduced location-space until the number of possible configurations is smaller than a user defined threshold (step 9 of Figure 3).

7. Compute E for all the remaining configurations to find the best one (exhaustive search; step 10 of Figure 3).

Results

The algorithm was tested on five high resolution crystal structures: Brookhaven Protein Data Bank (Bernstein et al., J. Mol. Biol. 1997; 112: 535-542) files: Bovine Pancreatic Trypsin Inhibitor (5PTI), RNAse-A (5RSA), Trypsin (1NTP) and carbonmonoxymyoglobin (2MB5) for which the neutron diffraction coordinates are available for proton positions, and phosphate-binding protein (1IXH) for which very high resolution results have been reported by X-rays. All hydrogen atoms were removed from the PDB files and the algorithm was activated to reconstruct their locations, assuming them to be in optimal positions in the crystal.

Each system was treated by two variations of the method:

1. Combined "ensemble-stochastic approach": Each system is divided into ensembles. Each ensemble is treated separately. All possible combinations in an ensemble are evaluated, and the one with the lowest energy is the result. In ensembles with a very large combinatorial demand the "stochastic approach" was activated to reduce the number of combinations to a number that could be exhaustively evaluated. The advantage in this approach is the short CPU time required for the calculation. As an example, the calculation on 3INS with its water layer by this method is interactive on a Silicon Graphics R10000 machine and takes about 4 minutes. However, this approach requires an approximation of distances between non trivial hydrogens and lone pairs in different ensembles, as described previously and the accuracy is somewhat reduced.

2. Pure "stochastic approach": The program is forced to treat the system as one huge ensemble. The algorithm reduces the number of combinations to a number that can be evaluated exhaustively. All hydrogens in the protein are added simultaneously in each combination-therefore no approximation is applied during the energy evaluation. This is important when there are minor energy differences between a few combinations and the accumulation of many long-range electrostatic interactions can add an important contribution to the final result. This approach has a larger CPU demand: The calculation on the same system takes about 15 minutes on a Silicon Graphics R10000 machine.

Given a system and an energy function, a test was devised to clarify whether the pure "stochastic approach" can find the global energy minimum out of a large number of possible combinations. To overcome the time limit for this type of calculation, an imaginary protein was constructed. It has 1186 amino acids, as illustrated in Figure 7, out of which 13 are serines (presented as CPK models) (13 segments) and 1173 glycines (0 segments). It has a globular shape with sizes 64Å *64Å *61Å . The serine hydroxyl oxygens were positioned to be at least 10Å apart. In this case, the interactions between the hydroxyls can be neglected, and each segment can be treated as a separate ensemble. All possible combinations in this ensemble may be evaluated to obtain the global minimum for the system. Thus, the pure "stochastic approach" was compared to the ensemble approach, which -in this unique case - is nearly equivalent to a full exhaustive evaluation. The stochastic search began with a total number of $5.02*10^{10}$ combinations and reached $2.7*10^{3}$ combinations after 204 iterations, which were then evaluated exhaustively. The ensemble method required only 1+4+15+12+10+11+2+10+6+12+5+8+11=107 calculations (the sum of positions of all

segments). The two methods yielded the exact same results for energy and for proton locations.

Five protein systems that have high resolution coordinates from a combination of X-rays and neutron diffraction have been analyzed. Only 5PTI, 5RSA and 2MB5 have many water molecules in their solvation shell, including proton positions.

Bovine pancreatic trypsin inhibitor (5PTI, 1.8Å resolution)

The structure of trypsin inhibitor was determined by joint X-ray (1.0Å resolution) and neutron diffraction (1.8Å resolution) (Wlodawer et al. J Mol. Biol. 1987;193:145–156). This PDB file contains 58 amino acid residues and coordinates for 63 water molecules. A 2.5Å water layer containing 54 water molecules was included in this calculation. A potassium and PO_4^{3-} ions from the PDB were also included in the calculation. The atoms in the side chains of residues GLU 7 and MET 52 were found to occupy two major sites. The *A* form was chosen for the calculation. Groups of rotatable atoms at a distance lower than 4.5Å were defined as one ensemble. The total was 21 ensembles and 256 possible locations.

The "combined ensemble-stochastic approach" was employed. In Table I, the number of possible combinations in an ensemble is the product of the number of combinations in each segment. The total number of combinations of ensemble 9 is the result of multiplying the number of positions for each segment, thus reaching 6,403,320. This ensemble was solved in a stochastic manner, while other ensembles were solved exhaustively. "Total energy" is the sum for all the ensembles. The lowest energy, for all the combinations in separate ensembles, was -121.0 Kcal/mole, while the highest energy was 2.9E+16 Kcal/mole. This high energy value is the result of "bumps" between rotatable hydrogens, which could not be eliminated at the preprocessing stage because only single proton bumps are tested in that stage.

The behavior of the system in the pure "stochastic approach" is demonstrated in Figures 8 and 9a. Figure 8 depicts ln(total number of possible combinations) vs. the iteration number. The initial number of combinations is $1.19*10^{30}$, of those, only 2690 remain for the exhaustive calculation after 443 iterations.

Figure 9a depicts the energy distribution in the 1st and 4th iterations. The x-axis does not hold the same energy values for all iterations: The average energy of the samples taken decreases in progressive iterations. Therefore, the samples are divided among 30 columns: lowest energy samples are in column 1, highest in column 30. The number of samples taken

in all iterations is constant. It can be seen that the algorithm eliminates energy bumps along the iterative process. Therefore, the energy distribution becomes more bell shaped along.

RNase-A (5RSA, 2.0Å resolution)

The structure of Ribonuclease A was determined by joint X-ray and neutron diffraction (2.0Å resolution) (Wlodawer et al., Acta. Crystallogr. B 1986; 42:379–387). This PDB file contains 124 amino acid residues, a PO_4^{3-} ion and coordinates of 128 water molecules. A 2.5Å water layer containing 90 water molecules was included in this calculation. The four histidine residues of 5RSA were retained in the calculation in their protonated form, as found in the PDB file.

Groups of rotatable atoms at a distance lower than 4.5Å were defined as one ensemble. A total number of 37 ensembles and 485 possible locations (Table II) was received. The "combined ensemble-stochastic approach" was employed. Ensembles 2, 7, 10, 29 that contained many combinations were solved in a stochastic manner, while other ensembles were solved exhaustively. The lowest energy, which is a sum of all lowest energy combinations in separate ensembles, was -60.8 Kcal/mole, while the highest energy was 261.3 Kcal/mole. Combinations with high energy values were excluded in early stages by a preprocessing "bump" calculation.

The behavior of the system in the pure "stochastic approach" is demonstrated in Figures 8 and 9b. The initial number of combinations is $1.76*10^{59}$, of those, only 2772 remain for the exhaustive calculation after 668 iterations.

Figure 9b depicts the energy distribution in the 1^{st} and 4^{th} iterations. Due to the absence of energy bumps, the energy distribution remains bell shaped during the minimization.

Myoglobin (2MB5)

The structure of myoglobin was determined by neutron diffraction (1.8Å resolution) (Cheng & Schoenborn, Acta Crystallogr. B 1990;46:195-208.). This PDB file contains 153 amino acid residues and coordinates for 89 water molecules (including their protons). It contains Protoporphyrin with Fe, an ammonium ion and a sulfate ion. All waters, ions and the Protoporphyrin moiety were included in the calculation. The HEM CO atoms are disordered. The *A* form was chosen for the calculation.

The "combined ensemble-stochastic approach" was employed, as illustrated in Table III. Groups of rotatable atoms at a distance lower than 4.5Å were defined as one ensemble. A total number of 43 ensembles was obtained.

The behavior of the system in the pure "stochastic approach" is demonstrated in Figures 8 and 9c. The initial number of combinations is $4.98*10^{52}$, of those, only 2400 remain for the exhaustive calculation after 552 iterations. Figure 9c depicts the energy distribution in the 1st and 4th.

Trypsin (1NTP, 1.8Å resolution)

The structure of trypsin was determined by neutron diffraction (1.8Å resolution)(Kossiakoff, Basic Life Sci 1984; 27:281-304). The enzyme is inhibited by a monoisopropylphosphoryl derivative, which was taken into account in the calculation. A calcium ion with a 2+ charge was added according to the indications in the PDB file and was positioned close to GLU 70, ASN 72, VAL 75 and GLU 80. This structure does not contain any water of crystallization. Groups of rotatable atoms at a distance lower than 4.5Å are defined as one ensemble.

Again, the "combined ensemble-stochastic approach" was employed. Table IV lists the total number of 33 ensembles with a minimal energy of 483.9Kcal/mole.

The behavior of the system in the pure "stochastic approach" is demonstrated in figures 6 and 7d. The initial number of combinations is $9.63*10^{10}$, of those, only 1152 remain for the exhaustive calculation after 14 iterations.

Phosphate-binding protein (1IXH, 0.98Å resolution)

The structure of Phosphate-binding protein has been determined by X-ray diffraction (Wang et al., Nat. Struct. Biol. 1997;4:519–522). The PDB file contains 321 amino acid residues. No water molecules' coordinates are reported. The protein is complexed with a PO_4 phosphate ion with a charge of -3. The ion was included in the calculations. This entry contains six disordered residues: Glu 1, Ser 3, Thr 162, Pro 216, Ser 234, Lys 245. The *A* form was chosen for all of them. The "combined ensemble-stochastic approach" was employed, as illustrated in Table V. Groups of rotatable atoms at a distance lower than 4.5Å were defined as one ensemble. A total number of 45 ensembles was obtained.

The behavior of the system in the pure "stochastic approach" is demonstrated in Figure 8. The initial number of combinations is $1.18*10^{21}$, of those, only 2400 remain for the

exhaustive calculation after 51 iterations.

Discussion

The five systems should be divided into two categories: The first are systems that lack experimental data for the coordinates of water molecules. Those systems are trypsin (1NTP) and the Phosphate-binding protein (1IXH). Figure 10 shows a Ribbon display of 1NTP and its polar residues. Many polar hydrogens should create hydrogen bonds to water molecules. However, no water coordinates are included in this PDB entry. The method of the present invention lacks, in this case, essential data for correct positioning of polar protons for residues on the protein's surface.

5RSA, 5PTI and 2MB5 are systems with much experimental data regarding water positions. Those are the three most important for this study, and a good algorithm is expected to yield accurate proton predictions for them.

The results of the methods for locating protons in biomolecular structures should be evaluated by a few criteria. First, the quality of the results should be examined in comparison to previously described methods as well as with respect to the ultimate goal, which is to achieve a negligible RMS for theoretical proton coordinates compared to experimental ones.

The "combined ensemble-stochastic approach" and pure "stochastic approach" results were compared to experimental, to a CVFF minimization using the MSI Discover/InsightII software package, to the method of Brunger and Karplus, and to that of Bass et al., as shown in Table VI. The CVFF minimization employed the "steepest descents" algorithm for the first 100 iterations, followed by conjugate gradients until convergence with a maximum derivative lower than 0.001 Kcal/Å was achieved.

The improvement with the methods of the present invention (ensemble-stochastic and the "pure stochastic") compared to positioning of protons by standard programs such as Insight(BIOSYM/Molecular Simulations. Discover 2.9.7 Force field simulations user guide 1995; Part 1; BIOSYM/Molecular Simulations. InsightII 95.0 Molecular Modeling System User Guide; 1995) with additional optimization by Discover/CVFF (BIOSYM/Molecular Simulations. Discover 2.9.7 Force field simulations user guide 1995; Part 1; BIOSYM/Molecular Simulations. InsightII 95.0 Molecular Modeling System User Guide; 1995) is clearly demonstrated. Self consistency algorithms, such as that of Brunger and Karplus (Proteins 1988;4:148-156) usually give better results than non specific methods. They are, however, less accurate than the method of the present invention. The method of the

present invention gives better results than Bass et al. in the more experimentally accurate 5RSA and 5PTI (see RMS values of Ser, try and water in Table VI) and similar results in the less accurate system, 1NTP.

The present invention has two additional improvements over Bass at al. First, there is no limit to the size of an ensemble, as systems can be treated as one huge ensemble (the "pure stochastic" approach) with some 92 segments (5RSA), while Bass et al. (Proteins 1992; 12:266-277) are limited to much smaller sizes. From Tables I-V it is clear that the close distance between rotatable protons in several regions of proteins, taken together with the number of options for positioning each proton, requires extremely long calculations if all options have to be considered. Since special attention must be paid to the need to locate protons in large molecules in a relatively short time, a stochastic method is better equipped to treat sizable molecules. Second, the method of the present invention is energy based, while Bass at al.(Proteins 1992;12:266–277) method is not.

No consistency was found with respect to improved prediction of a single residue type over others. This is also true for the other methods for locating protons. However, in some cases we find a correlation between the order of our RMS results for residue types and those of Bass et al. This may be linked to the spatial distribution of these residue types in each protein: some are closer to the core of the protein while others are closer to the protein's surface, and may be less accurate due to missing information about water positions.

One might expect a pure stochastic technique to drop some low energy solutions along the iterative calculation that excludes solutions, and therefore to yield less accurate results. It is remarkable to find how well it performs compared to the ensemble-stochastic method, which solves most of the systems in an exhaustive manner. The "imaginary protein" described in the results section was employed to compare the "pure" stochastic approach to an exhaustive search. Both techniques give the same minima out of $5.02*10^{10}$ possible combinations. This is a supplementary hint for the robustness of the "pure" stochastic approach as a tool for finding the global minimum.

One may gain information about the system's characteristics by inspecting the energy distribution charts (Figure 9). A bell shaped distribution in the first iterations indicates that there are no bumps between rotatable hydrogens. The "regular" bell shape of energy distributions for rotatable protons' positions, obtained after a few iterations, may be an expression of the proteins' density in the vicinity of those protons: a "dense" protein should increase the barriers for rotations. Thus, its energies should be skewed towards the high end

of the energy spectrum. The bell shape may be a demonstration of relative "free rotation" of those protons in a less dense surrounding.

Section 2: Location of Amino Acid Side Chains

The present invention is also particularly useful for solving the problem of correctly determining the locations of amino acid side chains within a protein. This specific implementation of the present invention solves a difficult problem, by enabling such locations to be determined with some accuracy, without undue assumptions but also without a combinatorial explosion.

The specific implementation of the present invention which is described in this section under "Methods" was also tested against other methods known in the art, as described under "Results". It should be noted that these methods and results are presented for the purposes of illustration only, and are not intended to be limiting in any way. The interpretation for these results is then discussed under "Discussion".

Methods

The search technique

The code uses a backbone dependent rotamer library. (Bower et al., J. Mol. Biol. 1997; 267: 1268-1282; Dunbrack & Karplus, Nat. Struct. Biol. 1994; 1: 334-340; Dunbrack & Karplus, J. Mol. Biol. 1993; 230: 543-574). For the purposes of testing only, and without any intention of being limiting, the August 1997 update of the rotamer library of Dunbrack & Karplus was used in the tests described below. A united atom model is employed (Weiner et al., J Amer. Chem. Soc. 1984; 106: 765-784). Energy is computed by equation 1 with the AMBER non bonding 12-6 Lennard-Jones and electrostatic energy terms, where $A_{i,j}$ is the repulsion parameter for the two (i, j) atoms, $B_{i,j}$ is their attractive polarizability parameter, q_i is the partial charge, $r_{i,j}$ is the distance between atoms, and ε is the dielectric constant. A distance dependent dielectric constant of $\varepsilon=r$ has been employed. V_n is the torsional potential barrier height for a torsion angle φ, n being the multiplicity and γ the phase factor. The potentials for V_n have been taken from the AMBER force field parameters. The non bonded energy is calculated for interactions with the backbone and with other residues' rotamers. The torsion energy term is calculated for all dihedral angles of each residue's rotamers. If the non bonded energy term exceeds the value of 10 Kcal/mole for a given pair of atoms, it is

truncated to 10 Kcal/mole.

$$(1)\quad E_{pot} = \sum_{i<j}\left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} + \frac{q_i q_j}{\varepsilon * r_{i,j}}\right) + \sum_{dihedrals}\frac{V_n}{2}[1+\cos(n\Phi-\gamma)] + \sum_{side-chains} -\ln\left(\frac{p_{rotamer}}{P_0}\right)$$

As suggested by Bower et al. (J. Mol. Biol. 1997; 267: 1268-1282) and implemented in the SCWRL algorithm, every rotamer is given a local energy based on its probability in the backbone-dependent rotamer library. Energies are taken from the probabilities of the backbone-dependent rotamer library, as $-\ln(p_{rotamer}/p_0)$, where p_0 is the probability of the most probable rotamer, and $p_{rotamer}$ is the probability of a given rotamer (assuming kT=1). The search strategy includes several steps:

(I) Steric clashes elimination stage and preliminary rotamer location: The input for the calculation are the backbone (N, C_α, C, O) coordinates of a protein with known structure. Those, together with φ and ψ angles of the backbone are used in order to create the initial placement of possible rotamers for each residue. Possible disulfide bonds between cysteine residues are calculated by the distance between sulfur atoms. All rotamers that clash with the backbone are excluded. If all rotamers of a residue clash with the backbone, the rotamer with the lowest "clash energy" remains. The algorithm treats single rotamers as part of the backbone, i.e. other rotamers that clash with those residues will also be excluded. The algorithm also searches for all side chain clashes between rotamer i of amino acid j and rotamer k of amino acid l. The algorithm excludes such pairs from being part of the solution, and therefore they are not sampled in the stochastic stage (*vide infra*).

(II) Stochastic stage: It is obvious that in the case of a large biological system such as a protein, a very large combinatorial problem results. In Hydrolase (1arb) (Tsunasawa et al., J. Biol. Chem. 1989; 264: 3832-3839), for example, there are $2.29*10^{105}$ alternative positioning options following step I. To reduce the size of the problem, the novel stochastic algorithm is employed. In the protein, the side chain rotamers in d_0 amino acids are unknown. For each amino acid there is usually more than one rotamer, but only one would give the lowest energy. Let $X_j=(x_{j1}, x_{j2}...x_{jd0})$ be a conformation of the protein which includes randomly picked rotamers for d_0 amino acids in a protein. For each conformation X_j, the energy $E_j=E(X_j)$ may be calculated according to the energy function described above. The objective is to find the conformation which minimizes E. Since it is impossible to evaluate all the alternative conformations due to the large number of combinations, the following steps are taken:

1. Sample at random n conformations out of the large population of combinations

$X_1=(x_{11}, x_{12}, \ldots, x_{1d0}), \ldots, X_n=(x_{n1}, x_{n2}, \ldots, x_{nd0})$, where x_{11} is a randomly picked rotamer for the first amino acid in the first conformation, and x_{n1} is a randomly picked rotamer for the same amino acid in the n^{th} conformation. We use n=1000 to create a large enough number of protein conformations, and compute the corresponding energy values: $E_1=E(X_1)$ to $E_n=E(X_n)$.

2. Construct the distribution F_E^n (n=10^3). F_E^n is the set of energies of all the n sampled conformations for the full protein. Define cutoff points H and L in F_E^n. H contains all variable values satisfying $E_i \geq F_E^n(1-\alpha)$, where $F_E^n(\alpha)$ is the αth percentile of F_E^n, while L contains all variable values satisfying $E_i \leq F_E^n(\alpha)$. The number of conformations in each of H and L is $n_0=n*\alpha$. When n=1000 conformations and α=0.01 (1%) for highest and lowest energy conformations, $n_0=\alpha*n=0.01*1000=10$ so L=10 and H=10. In other words, H stands for the 10 highest energy conformations, while L stands for the 10 conformations with the lowest energy.

3. Construct the vector h for all rotamer variables corresponding to the conformations in H. The vector h is the element-wise intersection of all the rotameric states in H, in the following manner: if all rotameric states in H share the same rotamer at component j (corresponding to x_{nj} of conformation X_n), then h_j=rotamer_number; otherwise, h_j=0 (no common rotamer for j in all high energy conformations.)

4. Construct the vector l for rotamer variables corresponding to the conformations in L. Unlike vector h, more than one rotamer may appear for each amino acid j up to a maximum of n_0 values in l_j. It is the union of all rotamers of component j that appear in the low energy conformations of L.

5. Compare h and l. If both h_j and l_j have a similar rotamer, it will remain as a viable rotameric state, because it contributes also to low energy values. However, if h_j does not correspond to any element of l_j, than the corresponding rotamer h_j will be evicted from subsequent iterations. If an amino acid has only one rotamer, it will not be evicted from subsequent iterations because it is the only remaining solution.

6. Repeat steps 1 to 4 for the reduced set of variables' values until the number of possible combinations of all variables is smaller than a user defined "end of stochastic stage criteria".

The value of α that is used to determine n_0 should be selected with care. If α is too large, no rotamers will be eliminated. If α is too small, an unjustified elimination of rotamers might occur. At best, α should be adjusted by the number of possible rotamers of each amino

acid, to allow an equal probability for the elimination of rotamers. In order to explain the determination of α, let us assume that each rotamer is not affected by interactions with any other amino acid in its environment. The α values for 2 to 29 possible rotamers of a single residue, that would lead to the correct rotamer elimination with a certainty > 99.983% are presented in Figure 14. Those values were calculated in the following manner. Given a residue with three rotamers, if we want to remove one rotamer with a certainty ($P_{correct}$) higher than 99.99%, the error probability (P_{error}) must be smaller than 0.01% (0.0001). For evicting erroneously a rotamer, it must first appear in all the high-energy conformations. In this case the probability is $(1/3)^{\alpha}$ In addition, this rotamer must not appear in any low energy conformation. In this case the probability is $(2/3)^{\alpha}$ The total error probability is $P_{error}=(1/3)^{\alpha}(2/3)^{\alpha}$. Thus, one may tune the calculation to nearly 100% confidence by employing the general formula $P_{error} = \left(\frac{1}{m}\right)^{\alpha}\left(\frac{m-1}{m}\right)^{\alpha}$ where m is the number of variable values (rotamers). When m=1 (there is one rotamer) P_{error}=0. Assigning a value of P_{error}=0.0001 and solving the equation leads to a value of α=6.12. When α is very large, P_{error}=0, but the odds of evicting any variable value are very low. Thus, the α values are preferably employed from Figure 14, which allow eviction of variable values, with $P_{correct}$=99.983%-99.9988 %.

(III) end of search: Once there are less than M combinations remaining (M~10^5), an exhaustive search is conducted to yield the N lowest energy conformers of the protein.

Results

The stochastic algorithm is applied to 10 proteins of various sizes (46 to 263 residues), and complexity ($1.04*10^{14}$ to $2.29*10^{105}$ possible combinations after elimination of rotamers that clash with the backbone), that were chosen to cover a range of protein fold families. Out of these 10 proteins, 6 (46-68 residues) were also selected by Leach & Lemon (Proteins 1998; 33: 227-239) employing the DEE/A* algorithm, and those serve to compare between the stochastic and the DEE/A* algorithms. These proteins are: Crambin (PDB entry 1crn) (Teeter et al., J Mol Biol. 1993; 230: 292-311), Ribosomal protein (1ctf) (Leijonmarck & Liljas, J Mol Biol. 1987; 195: 555-579), Complement control protein (1hcc) (Norman et al., J Mol Biol. 1991; 219:717-725), Ovomucoid third domain (2ovo) (Empie & Laskowski, Biochemistry 1982; 21: 2274-2284), Erabutoxin B (3ebx) (Smith et al., Acta Crystallogr A. 1988; 44:357-368), and Rubredoxin (5rxn) (Watenpaugh et al., J Mol Biol. 1980; 138:

615-633). The remaining proteins selected were larger (129-263 residues), with high resolution X-ray structures (resolution < 1.5Å, R factor < 0.17): Lysozyme (2ihl), Ribosomal protein (1whi) (Davies et al., Structure 1996; 4:55-66) Endonuclease (2end) (Morikawa et al., Science 1992;256:523-526) and Hydrolase (1arb) (Tsunasawa et al., J. Biol. Chem. 1989; 264:3832-3839). Table VII summarizes the results of applying the stochastic algorithm to the 10 proteins. For each protein, the number of combinations (following initial exclusion of rotamers that clash with the backbone) is presented, with several values for single conformations (the global minimum for each protein) and average values for a "population" of 1000 low energy conformations. The best possible RMS is depicted for each protein. Finally, the average energy gap of those 1000 conformations (without weighting) is presented. RMS were calculated for side-chain atoms (excluding C_β) of the global energy minimum conformation compared to the X-ray conformation. The RMS range is 1.32-2.60 for the global minimum. Average RMS values for the 1000 low energy conformers are somewhat larger than for the global minimum, but for each protein, conformations that are higher in energy than the global minimum are found, that have a lower RMS than that minimum. The range of energy values for the 1000 lowest energy conformers is up to 5.52 Kcal/mole above the global minimum. The average energy gap of the 1000 lowest energy conformers from the global minimum is always small (2.20 Kcal/mole for all the proteins).

A test of the search method's validity

In order to test the efficiency of the stochastic search, and in view of the values reported in table VII, a number of questions were raised. The first question is whether the stochastic search achieves the results that could be obtained by an exhaustive search, given a specific rotamer library. The second questions is whether such a search can identify the crystallographic structure of a protein if the rotamer library includes the original X-ray rotamers.

The first question requires a test of a relatively small protein, in which such an exhaustive search may be carried out. Given the constraints of the energy function and the rotamer library, our stochastic algorithm was imposed to find the lowest energy combinations in a test protein and compare them to the results of an exhaustive search. The protein selected was crambin (PL form), Brookhaven Protein Data Bank (Bernstein et al., J. Mol. Biol. 1997; 112: 535-542) file 1cnr (Teeter et al., J Mol Biol. 1993; 230: 292-311) which is a high quality X-ray structure (1.05Å resolution, *R* factor=0.105). It is large enough to constitute a

reliable test case, but not too large, which would require long computations in an exhaustive search. The entry contains 46 amino acid residues (see Figure 11) and coordinates for an ethanol molecule. There are 8 disordered residues (Thr 1, Thr 2, Ile 7, Val 8, Arg 10, Asn 12, Ile 34, Thr 39). In order to evaluate this protein in a reasonable time period, Arg 10 (the A form in this disordered residue), Arg 17, Glu 23, Ile 33 and Ile 35 were kept fixed in their original positions. The initial number of combinations (following the eliminative step of steric clashes) was $6.79*10^8$. In Figure 12, the results of the stochastic and exhaustive searches for a range of N low energy conformations are compared. Energy values and % difference between the two searches are depicted for each of the 10,000 conformations. The 485 lowest energy conformations of this protein are found to be exactly the same by the stochastic and the exhaustive searches. For a large number of conformations, the difference is minor. It may be seen that the energy of conformer no. 10,000 is 4.71 Kcal/mole higher than the global minimum in the exhaustive search, and 4.80 Kcal/mole in the stochastic one. Thus, the difference between these searches for that conformer is only 0.56%. This test was repeated with three different seed numbers for the random numbers generating function (100000, 200000 and 300000) with similar results.

For testing its ability to reproduce the X-ray coordinates, the stochastic algorithm was employed with an extended rotamer library to which the crystal rotamers of 1cnr were introduced. No residues were fixed during this search. Energies were computed by equation 1 without the probability term, which is not available for the crystal coordinates. The following residues were not included: four Gly (no side chain), five Ala (only one possible rotamer) and six Cys (no rotamers because all of them form S-S bonds). Therefore, out of 46 amino acids in the sequence, 31 remained for this comparison. The energy of the protein in its crystal structure coordinates was 3.41 Kcal/mole higher than the global minimum found by the stochastic algorithm. In 20% of the residues (6 amino acids: Ser 6, Val 8, Thr 21, Tyr 29, Ile 33, Ile 34) a full superimposition of the search results over the X-ray ones was obtained. In 58% of the residues (18 amino acids) a high quality superimposition was found: the absolute angle deviation of all torsion angles was found to be less than 40°. So, the extended rotamer library, for which the RMS should be 0.0, located correctly some 80% of the side chains. For example, atoms CG of Leu 18 were 0.18 apart, and CG atoms of Asn 14 were 0.23Å apart. The RMS value between the global minimum structure and the crystal structure for side-chain atoms (excluding C_β) was 1.16.

To test the limitations of the original rotamer library (with no crystallographic

rotamers), each rotamer was located as close as possible to the relevant side chain in the crystal structure. The RMS value obtained was 1.15. The RMS value between the global energy minimum in the stochastic search and the crystal structure was found to be 1.97.

Comparison of the algorithm to results from X-rays, NMR and MD

The conformational space of E. coli ribonuclease HI was explored with the method of the present invention, for comparing the results for the lowest energy 1000 conformers to experimental and theoretical methods that offer an insight into the conformations that each side chain may adopt under different conditions: X-ray crystallography, NMR, and MD.

An ensemble of 8 NMR structures (PDB entry 1rch) was reported based on distance restrictions from experiment. Philippopoulos & Lim (Proteins 1999; 36: 87-110) compared an extended set of NMR results to the high-resolution (2rn2, 1.48Å) crystal structure (Katayanagi et al., J Mol Biol. 1992;223:1029-1052), to the lower resolution (1rnh, 2.05Å) crystal structure (Yang et al., Science 1990; 249: 1398-1401) and to their own MD simulations. NMR and MD simulations yield few results for each torsion angle, and resulting conformations were classified as ensembles. Each ensemble is represented by the mean value of its dihedral angles and an order parameter, S, (Hyberts et al., Protein Sci. 1992; 1:736-751), which expresses the deviation of each dihedral angle from its mean value. The S parameter of each dihedral angle in each residue is calculated in turn across the ensemble. The order parameter $S(\alpha_i)$ for an angle α_i of residue i (where $\alpha=\varphi$, ψ, χ_1 ,$\chi_2$etc) is defined as: $S(\alpha_i)=1/N*|\Sigma_{\varphi=1}{}^{N}\alpha_i{}^{\varphi}|$, where N is the total number of structures in the ensemble, α_i^j (j=1, ..., N) is a 2D unit vector with phase equal to the dihedral angle α_i, i represents the residue number, and j stands for the number of ensemble number. If the angle is the same in all structures than S has a value of 1, whereas a value of S much smaller than 1 indicates a disordered region of the structure. Philippopoulos & Lim limited their classification to an S value greater than 0.8.

The stochastic algorithm was employed on the backbone of 2rn2, which has a higher X-ray precision. The calculation started with $1.61*10^{87}$ possible combinations. The algorithm has been employed to refine the conformational space into $1.3*10^{33}$ best conformations by evicting high energy conformers, which leaves enough conformers to evaluate the conformational flexibility of the protein.

Table VIII contains a comparison between the stochastic algorithm, and the results of X-ray crystallography, NMR and MD. This table focuses on residues adopting highly

probable conformations according to the following assumptions: In some cases torsion angles assumed a single conformation in the MD ensemble and multiple conformations in the NMR ensemble, while in others the reverse was obtained. We assume a high probability for an experimental rotamer if it obeys one or more of the following rules: (1) It appears in the high resolution crystal structure (2rn2). (2) It is found in at least two out of the three: low resolution crystal structure (1rnh), a NMR model and the MD simulation. A "hit" was considered to be any result of the stochastic algorithm, which has a fluctuation of up to $\pm 30°$ from the "correct" conformer. Each such hit is marked by a "+" in the table. In some cases angles such as $\chi 1$ of M 47 are presented by a single rotamer in the table, and marked by "(+)". Such angles have additional values that do not obey the above two rules. Those other angles are considered to have low probability, and do not appear in table VIII. Out of 115 dihedral angles in table VIII, 7 angles are missing from the rotamer library (see Figure 13A), and two other angles deviate by $\sim 40^0$, and therefore were not included in our evaluation as "hits". Thus, we may expect a maximum of 106 "hits", in comparison to X-rays, NMR and MD. The stochastic algorithm predicts correctly 87 angles (see Figure 13B), which is 82%.

Comparison of the algorithm to the DEE/A* algorithm

Leach & Lemon (Proteins 1998; 33: 227-239) explored the conformational space with the DEE/A* algorithm on a set of 8 proteins chosen to cover a range of protein fold families. The method of the present invention was then employed on 6 of those proteins (1crn, 1ctf, 1hcc, 2ovo, 3ebx, 5rxn). Snake venom neurotoxin (1nxb) (Tsernoglou et al., Mol Pharmacol. 1978; 14:710-716) was excluded due to an unknown residue type (residue 59). Bovine pancreatic trypsin inhibitor (5pti) (Wlodawer et al. J Mol. Biol. 1987; 193: 145-156) was excluded due to the occupation of two major sites by residues Glu 7 and Met 52. Leach & Lemon also explored the effects of "united" atom and "all atom" models with "standard" and "reduced" electrostatic representations. Unfortunately, they did not report RMS values of each system separately, but only an average value for all 8 systems in each search method. Table IX contains a comparison between the stochastic method and DEE/A*. The maximal number of combinations solved by the stochastic algorithm is $2.29*10^{105}$, while DEE/A* reached only $2.48*10^{34}$ combinations. The largest protein system solved by the stochastic algorithm is 263 amino acids, while DEE/A* solved a maximum of 68 residues. In order to assess the correctness of models, the average RMS for side-chain atoms (excluding C_β) of the predicted conformation and that of the X-ray conformation was then calculated. The best

possible RMS for the current rotamer library is depicted. The stochastic algorithm's RMS values range was found to be 1.32-2.48, with an average of 2.07 in the same systems that were also calculated by DEE/A*. The best possible average RMS with our rotamer library is 1.18 for all proteins in our test case. Leach & Lemon reported average RMS values from 1.77 to 1.92 depending on the atom model and rotamer library, with a best possible RMS for rotamer library that ranges between 0.75-0.83. On the larger systems, for which DEE/A* could not be employed due to combinatorial explosion, the stochastic algorithm found an average RMS ranging from 2.22 to 2.60 with an average of 2.40. The best possible RMS for the rotamer library was 1.23.

Discussion

The previous description concerns the application of a novel stochastic search technique to explore the conformational space of proteins' side chains. It is an extension and refinement of the above example in the previous section for searching the positions of polar protons in proteins. The algorithm successfully explores the conformational space of various sizes of proteins and can deal with a large number of combinations after eliminating rotamers that clash with the backbone.

The robustness of the stochastic algorithm in handling complex combinatorial searches is clearly demonstrated in Tables VII and IX. Comparing it to an exhaustive search (Figure 12) proves the reliability of the stochastic algorithm in finding increasing amounts of lowest energy conformations. For 485 low energy conformations of this protein, no difference between the stochastic and exhaustive search was found. When the limit of 10,000 lowest energy conformations is approached, a minor deviation of 0.56% has been detected. Since this large number of conformations reaches, at its maximum, an energy gap of 4.71 Kcal/mole above the global minimum, this population includes the prevailing contributors to molecular properties according to the Boltzmann distribution. They represent the major contributions to the molecular partition function, which may be employed toward the computation of conformational entropy.

With no difference between the stochastic and an exhaustive search for the population of low energy conformations, another issue is the comparison of our search results to experiments. This has been presented in table VII, by comparing our global minimum to the crystallographic results of 10 proteins and in table VIII, by comparing our low energy populations to the detailed results of X-rays, NMR and MD for a single protein. The RMS

values of the global minima for the 10 proteins are strongly affected by the rotamer library, but not entirely: the energy expression is limited for reproducing the structure. Including the original crystallographic rotamers in that library has proved this point. Even in that case, only ~80% of the residues were calculated with high accuracy. The energy of 1cnr crystal coordinates was found to be 3.41 Kcal/mole higher than the global minimum found by our stochastic algorithm. However, the RMS value for that structure is only 1.16. Without crystallographic rotamers, the stochastic search in 1cnr results in a RMS of 1.97. The limitations imposed by the rotamer library are expressed in table VII by the column that presents the best possible RMS of this library for each of the proteins. These values contribute 50-75% of the error in the RMS values for the global energy conformations. It may be seen from table VII that the global minimum energy conformation is not necessarily the one with lowest RMS value. There are higher energy conformers whose structure is closer to the results of X-rays.

Table VIII contains 106 angles of E. coli ribonuclease HI which was expected to be detected by comparing to X-rays, NMR, or MD. The algorithm detected correctly 87 angles, which are 82% of the total. Part of the deviation from 100% accuracy may be due to the quality of the rotamer library, but a greater part is due to the energy function. Mendes et al. (Proteins 1999; 37: 530-543) presented a rotamer as a continuous ensemble of conformations that cluster around the classic rigid rotamer. Such a technique may increase the rotamer library's efficiency. The results (RMS of 1.16 with crystallographic rotamers, 1.97 without) support the claim that a larger rotamer library does not guarantee a dramatic improvement in RMS values (Proteins 1992;14: 213-223; J. Mol. Biol. 1994; 235: 1088-1097; Tanimura et al., Protein Sci. 1994; 3: 2358-2365;Vasquez, Biopolymers 1995; 36: 53-70).

Currently there are four main methods to study the conformational space of a given protein: X-ray crystallography, NMR, MD, and rotamer library based methods. X-ray crystallography usually suggests a single structure, which might be biased toward specific conformational substates in the crystal (Brunger, Nat. Struct. Biol. 1997; 4 suppl: 862 -- 865). Observing different conformations may be possible only at the highest resolution. The advantage of our algorithm is straightforward: it extends the single conformation into a family of viable conformations.

Unlike X-ray crystallography, NMR suggests alternative conformations by deciphering the 2D and 3D coupling maps. NMR does not teach us about the shape of the energy minima in the potential energy surface. NMR of proteins is a long and tedious

experiment limited by the time scale of conformational variations, especially in large proteins. In this case, the method of the present invention may be an additional tool for suggesting alternative conformations. When NMR structures are available, the method of the present invention may be employed to extend this information by allowing the determination of the conformations' energy weights, thus enabling an assessment of their contribution to the overall population at equilibrium.

MD simulations require extensive CPU time scales for biomolecules, which prohibits the full exploration of the conformational space. MD suggests conformations that may not be detected by NMR or by X-ray crystallography. MD time scales and barrier crossing ability are not yet reliable enough for detecting the global minimum or the population of lowest energy conformations in large biomolecules. The reliability of our stochastic algorithm in finding both has been demonstrated in this paper. However, while MD trajectories imply a mechanism of conformational interconversions, the stochastic approach concentrates on products and not on pathways.

Dill and Chan (Nature Struct. Biol. 1997; 4: 10-19; Chan & Dill, Proteins 1998, 30, 2-33) declared that the native state of a given protein corresponds to the global minimum in free energy, which is not necessarily the global minimum potential energy. Thus, an algorithm for adding side chains should yield most of the lowest energy conformations, to enable entropy evaluation. Currently, the method of the present invention meets this demand. In the "mean field" approximation, each side chain "feels" an average of all possible conformations of its neighbors. The conformational entropy is then estimated from the side chain probability of a given possible location (Vasquez, Biopolymers 1995; 36: 53-70; Koehl & Delarue, Nat. Struct. Biol. 1995; 2: 163-170; Koehl, & Delarue, Curr. Opin. Struct. Biol. 1996 ;6:222-226). The present stochastic search offers, in addition to finding the global minimum, the next N best solutions for rotamers in large proteins without any mean field approximation and is unique in that sense. It may thus be employed for studying thermodynamic properties of complex molecular systems. The stochastic algorithm can treat more than 250 residues (the maximum at this stage is 2.29^{105} combinations). The DEE/A* algorithm treated a maximum of 68 residues and the maximal number of combinations (before backbone clash exclusion) was 10^{44}. Following the application of the DEE algorithm, the size of the remaining space to be explored by the A* algorithm may be reduced to a maximum of 10^{21}.

The quality of the method of the present invention, with its energy expression and a

backbone dependent rotamer library, is compared to the results of the combined DEE/A* algorithm (Leach & Lemon, Proteins 1998; 33: 227-239), with a different energy expression and with two different libraries. A comparison of each technique to experiment by RMS is limited, because it is affected by the rotamer library : A RMS value of 2.0 with a rotamer library whose lowest RMS value for a protein is 1.9 reflects a better search technique than one with a RMS value of 1.5 obtained from a library whose optimal RMS is 0.1. The RMS values should be compared to the optimal RMS value that could be achieved within the constraints for the rotamer library. In Table IX one may see the correlation between the best possible RMS values for the library and the RMS of global energy conformation. This fact may explain the difference between these results and Leach & Lemon' s results. Another advantage is in the ability to employ the stochastic algorithm in a "stand alone" mode without any preprocessing algorithm (such as DEE in the case of the A* algorithm). The A* algorithm requires a good estimate of the cost dependence to reach a goal node. This might be difficult to achieve because interactions between residues that were not yet assigned to any position cannot be easily computed. One should also note that the numbers of combinations presented in tables VII and IX for the stochastic algorithm refer to possible numbers of combinations that remain after evicting rotamers that clash with the backbone. Hence, the real number of possible combinations is much higher.

Section 3: Prediction of Loop Structure in Proteins

The present invention is also particularly useful for solving the problem of correctly predicting the structure of loops within a protein. This specific implementation of the present invention solves a difficult problem, by enabling such predictions to be determined with some accuracy, without undue assumptions but also without a combinatorial explosion.

The specific implementation of the present invention which is described in this section under "Methods" was also tested against other methods known in the art, as described under "Results and Discussion". It should be noted that these methods and results are presented for the purposes of illustration only, and are not intended to be limiting in any way. The interpretation for these results is then discussed.

Methods

The construction of loops may be achieved by several strategies. Most of them employ standard bonds and bond angles, while varying dihedral angles only. This particular

implementation of the method of the present invention follows this general path, while deviating from it in several steps.

Geometric premises

Figure 15 depicts an example of 6 residues (0-5). Residues 0 and 5 are in the invariable part of the protein. A search is performed for the conformations of residues 1-4. The loop is constructed simultaneously from both the N and C-termini (Moult & James, Proteins 1986; 1: 146-163) and the loop closure is tested between residues 2 and 3. Such a construction strategy reduces the accumulation of errors: when one constructs the loop by dihedrals from one terminal toward the other, an inaccuracy in the first residues leads to an increasing amount of deviations in further residues.

Figure 16 depicts the dihedral angles definition for a given residue: ψ of a residue n, in the construction strategy, is the ψ of the previous residue toward the N-terminal. The thought behind such a definition is that both φ_n and ψ_n define the location of N and C atoms in residue n. When constructing the loop from the N terminus (starting from residue 1 in Figure 15), the nitrogen of residue 1, the first to be predicted, should be located according to the ψ angle of the former residue. The exemplary method of the present invention, as described below, assumes a trans (180°) structure for Cα-C-N-Cα. Thus, in residue 1, Cα is located according to this premise. The carbonyl carbon of residue 1 is located according to φ_1, which is extracted from the search (*vide infra*). The nitrogen of residue 2 is located according to ψ_2 (which is regularly defined as ψ_1) and so on. When constructing the loop from the C terminus, the carbonyl carbon of residue 4 is located by φ_5. Cα of residue 4 is located at a 180° to the Cα of residue 5. The N of residue 4 is located according to ψ_5. Likewise, residue 3 is located on the basis of φ_4 and ψ_4 as defined in Figure 16. Thus, the values of φ_3 and ψ_3 are not required.

Assigning residues' (φ ;ψ) possible angles

The odds of finding, in the structural database, entire peptide sequences with lengths of more than 6 residues are prohibitively poor (Oliva et al., J. Mol. Biol. 1997; 266: 814-830). Therefore, the method of the present invention employs a search for segments of 3 overlapping residues of each loop in SWISS-PROT (Bairoch & Apweiler, Nucleic Acids Res. 2000; 28: 45-48). Given a protein with a sequence ...ACGDEIL..., where 'A' is residue 0 from Figure 15, and CGDE is the loop, the method of the present invention searches for

ACG, CGD, GDE, DEI and EIL segments. The Brookhaven Protein Data Bank (Bernstein et al., J. Mol. Biol. 1997; 112: 535-542) is explored for all (φ ;ψ) angles of the relevant residues in the segments detected by the SWISS-PROT search. The search is conducted only for the second and third residues in each triplet, so that any φ ;ψ combination found must be associated with the order of the loop sequence. Such a search yields multiple allowed conformations, including rare ones and may result in a few hundred pairs of φ ;ψ angles for a given residue. Those are kept as a database for subsequent processing. If both φ ;ψ angles differ by less than 2° from another pair of the same residues, they are discarded from the database. Values of dihedral angle pairs for any protein that was explored, were eliminated from the database for testing this particular protein, in order to avoid any bias.

Exploring the conformational space with the stochastic algorithm

It is obvious that in the case of medium and large loops, a large combinatorial problem results. For example, the number of combinations in the second loop of bacteriorhodopsin complex at 1.55Å resolution, (Luecke et al., J. Mol. Biol. 1999; 291: 899-911) (Brookhaven file 1c3w.pdb) is $5.5*10^{28}$. Only a portion of the database conformations may close loops that obey the geometric criteria. To reduce the size of the problem, the method of the present invention is employed. Given a loop with d_0 unknown angle pairs, only a small part would participate in lowest cōst function (*vide infra*) possible structures. Let $X_j=(x_{j1}, x_{j2}...x_{jd0})$ be a conformation of the loop which includes randomly picked $((\varphi_{j1};\psi_{j1}), (\varphi_{j2};\psi_{j2}), ... ,(\varphi_{jd0};\psi_{jd0}))$ angles. For each conformation X_j, the cost function $C_j=C(X_j)$ may be calculated. The objective is to find all conformations which minimize C. The method was described previously in detail, in the previous two sections. Briefly, the following steps are followed:

1. Randomly pick a value for each pair of angles: the total constitutes a conformation of the full loop.

2. Employ the "cost function" to calculate the value of the current conformation.

3. Continue to calculate the value for n such conformations, each with all its variables' values picked randomly.

4. Construct a histogram of the distribution of values for all sampled energies (n~1000).

5. Compare all variable components that contribute to a portion α (α=number of conformations) of the full histogram, at its high-end region.

6. Evict components that contribute to all highest value conformations, but not to any lowest value ones.

7. Repeat the process iteratively until remaining combinations can be evaluated exhaustively.

At the end of this stage, many conformations remain which obey the geometric loop closure criteria, but are not necessarily clash free. Therefore, in the next stage side chains are added, and the loops are evaluated by energy criteria. In various tests of this algorithm, at the end of this stage, between 10^2-10^5 conformations were retained for further processing.

The scoring function in the stochastic stage

The purpose of the stochastic stage is to generate a population of loops that could potentially close. By employing equation 2, loops which remain open are evicted. The method of the present invention explores the conformational space using the cost function in equation 2.

$$C = \left| \sum_{i=1}^{4} r_i^{predicted} - r_i^{experimental} \right| \qquad (2)$$

where the distances d_i are shown in Figure 15. They are calculated following the positioning of the last connecting residues from the N and C terminals. Values of $r_i^{experimental}$ are standard ones such as N-C bond length(Shenkin et al., Biopolymers 1987; 26: 2053-85).

Scoring the loops

Once there are less than M combinations remaining (M~10^2-10^5), an exhaustive search is activated to yield the N lowest energy conformers of the loop. In order to score the energy of remaining loops, their side chains were added, employing a recently updated version of a backbone dependent rotamer library (Dunbrack & Karplus, J. Mol. Biol. 1993; 230: 543-574; Dunbrack & Karplus, Nat. Struct. Biol. 1994; 1: 334-340). For the atoms, a united atom model was employed (Weiner et al., J Amer. Chem. Soc. 1984; 106: 765-784). Backbone N-H and polar hydrogens of side chains are represented explicitly. AMBER (Weiner & Kollman, J. Comp. Chem. 1981; 2: 287-303; Weiner et al., J Amer. Chem. Soc. 1984; 106: 765-784) bonding and non bonding energy terms were employed (equation 3) with a distance dependent dielectric constant of $\varepsilon=2r$. The non bonding energy is calculated for interactions with the backbone and with other residues' rotamers. A 12-10 potential for hydrogen bonds, between all polar hydrogens and possible acceptors was employed. In order

not to lose solutions that could be satisfactory but present some VdW clashes, the Lennard-Jones repulsion energy was truncated at a value of 30 Kcal/mole for a given pair of atoms.

$$(3)\ E = \sum_{non-bonding}\left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} + \frac{q_i q_j}{\varepsilon * r_{i,j}}\right) * p_i * p_j + \sum_{h-bond}\left(\frac{C_{i,j}}{r_{i,j}^{12}} - \frac{D_{i,j}}{r_{i,j}^{10}}\right)$$

The function was employed in a mean field form. As suggested by Bower et al. (J. Mol. Biol. 1997; 267: 1268-1282) and implemented in the SCWRL algorithm, every rotamer is given a probability in the backbone-dependent rotamer library. The energy of interaction from equation 3 is multiplied by the probability assigned from the relevant rotamer (p), where the sum of rotamer probabilities is 1 for each residue. Rotamer-rotamer interactions, rotamer backbone interactions, and backbone-backbone interactions were all considered. A subset of residues that have at least a single atom at a distance of 10Å from the native loop was included as a "template". When atoms in equation 3 are from the backbone their probability is p=1. The bonding energy terms included stretching (equation 4), bending (equation 5) and torsion energies (equation 6). The stretching energy (equation 3) is calculated between the carbonyl carbon and the nitrogen that close the loop (d_1 between residues 2 and 3 in figure 15)

$$(4)\ E_{streching} = \sum_{bonds} k_b (r - r_0)^2$$

The "k_b" parameter controls the stiffness of the bond spring, while r_o defines its equilibrium length. A value of k_b=100 was assigned in order to soften the energy function. The bending energy is calculated as follows (equation 5)

$$(5)\ E_{bending} = \sum_{angles} k_\theta (\theta - \theta_0)^2$$

The "k_θ" parameter controls the stiffness of the angle spring, while θ_o defines its equilibrium angle. Unique parameters for angle bending are assigned to each bonded triplet of atoms based on their types. Two triplets were employed. The first was the Cα-N-C (d_2 in figure 15), where Cα is part of the previous residue. The second triplet included Cα-C -N, where Cα and C are part of the previous residue (d_3 in Figure 15).

The torsion energy is modeled by a periodic function (equation 6):

$$(6)\ E_{torsion} = \sum_{torsions} A[1 + \cos(n\tau - \phi)]$$

The "A" parameter controls the amplitude of the curve, the n parameter controls its periodicity, ϕ shifts the entire curve along the rotation angle axis (τ). Unique parameters for torsional rotation are assigned to each bonded quartet of atoms based on their types. The torsion energy of the angle between Cα-N-C- Cα atoms (d_4 in Figure 15), i.e. the energy "price" of deviation from a planar (180°) amide bond, is then calculated.

All results were evaluated by comparing to the loop coordinates from high resolution X-ray crystallography, by applying the coordinate root mean square deviation algorithm (cRMS). Such comparisons have been done only for N, Cα, and C of the backbone, in order to allow comparisons to other methods (van Vlijmen and Karplus, J. Mol. Biol. 1997; 267: 975-1001) and Deane & Blundell (Proteins 2000; 40: 135- 144)).

Results and Discussion

The above tests were intended to verify whether the novel stochastic search method may be applicable also to loop construction and whether it may be employed for the reconstruction of structurally known loops of varying size. The example used was a transmembrane protein. The only extensive experimental example is bacteriorhodopsin, which contains 7 transmembrane helices and was recently studied by high resolution crystallography (Luecke et al., J. Mol. Biol. 1999; 291: 899-911). The search was applied to this structure (X-rays results at 1.55Å resolution, PDB file 1c3w). The six loops of bacteriorhodopsin are listed in Table X. Loops 3 (CD, intracellular) and 4 (DE, extracellular) contain 2 and 1 residues respectively, and are not interesting test cases. In loop 5 (EF, intracellular) the coordinates are not included in the entry, thus the quality of results cannot be clearly assessed. The remaining loops: 1 (AB, intracellular) , 2 (BC, extracellular) and 6 (FG, extracellular) are attractive test cases and range from 4 to 16 residues. In order to avoid bias, the 1c3w.pdb entry was not included for creating the residues' (φ ;ψ) angle database that is employed for the stochastic search. The RMS values ranged between 0.28-2.46 (table XI), with an average value of 1.35. It is encouraging to see that the algorithm can yield on the one hand a very low RMS in the small loop and a good RMS value in the case of the very large loop.

A comparison was subsequently attempted for the efficiency of the stochastic loop prediction for globular proteins, by comparing it to the recent report of Deane & Blundell (Proteins 2000; 40: 135- 144) and to that of van Vlijmen and Karplus (J. Mol. Biol. 1997; 267: 975-1001). Deane & Blundell employed an ab initio loop construction method. Their

algorithm selects polypeptide fragments from a computer-generated database. Each fragment is defined by a representative set of eight (φ ;ψ) pairs. This fragment set is scored and sorted using a RMS fit to the anchor regions and a knowledge-based energy function. Van Vlijmen and Karplus employed a search on a database composed of 130 loops from 21 proteins. The best loops among the large number of candidates was determined by a CHARMM (Gunsteren & Karplus, J. Comp. Chem., 1980; 1: 266-274) non-bonded energy function (without electrostatics) applied to the backbone and C(β) atoms. The method of the present invention was tested on the longest seven of their eleven example loops. Table XII compares our results to the results reported by the two methods. The Average RMS values were 1.86 with a range of 1.06-2.99 in comparison to an average of 2.3 (0.3-5.2) in the case of Vlijmen and Karplus and an average of 2.1 (1.3-3.2) in the case of Deane & Blundell. These lower average RMS values clearly demonstrate the quality of the method of the present invention.

In addition, the algorithm supplies a large group of low-energy loop conformations, which may be further employed for evaluating loop properties such as flexibility, as well as for comparing, in case of reconstructing known loops from PDB, to the loop's temperature factors from crystallography.

Several basic questions were raised during this work. The first one concerned the accuracy of the approximation of employing standard bond lengths and angles. For that aim, the method of the present invention was employed on the first loop of bacteriorhodopsin (*vide infra*). The RMS value between predicted and experimental backbones was 0.280. The real experimental dihedral angles were added to the angles' database, and the rest of the dihedral angles were deleted. Hence, the only option for the method of the present invention was to construct the system according to the experimental dihedral angles. If the rest of the angles and bond lengths were similar to the experimental one, one might expect to obtain a RMS value of 0. However a RMS value of 0.204 resulted. It indicates that such an approximation has a minor but not negligible effect. One must take that into consideration, especially when building large loops where the accumulation of errors might skew the results.

The second question concerned the accuracy of approximation of evicting φ ;ψ angle pairs that differ by less than 2° from another pair of the same residues (for both angles). The previous test was repeated with a slight change: all the experimental dihedral angles were increased in 2°. Surprisingly, a RMS value of 0.198 resulted. Repeating the same test with a 2° decrease for all dihedral angles resulted in a RMS value of 0.220. With such minor

differences, the approximation can be shown to be appropriate.

Global optimization of a loop structure is a difficult task due to a strong dependence among variables: modifying a single φ or ψ angle might induce a dramatic conformational change in the entire loop. Thus, the question raised concerned whether the method of the present invention, which successfully located protons and side chains, could generate the population of loops that obey the geometric criteria defined in equation 2. Again the Bacteriorhodopsin first loop was employed. It is not prohibitively large for an exhaustive search, and on the other hand it still poses a combinatorial challenge. The 10,000 "lowest cost function" conformations for equation 2 are depicted in Figure 17. Both searches began from 54,330,000 combinations. A same global minimum was achieved by the two search techniques. The 66 first conformations were identical in their cost value (RMS). The worst error (RMS, in the 2018th solution) was 3.36 % with a cost value difference of 0.006721Å . This test demonstrates the ability of the method of the present invention to search effectively and to obtain significant results for small and large loops. These results strongly support the robustness of the method of the present invention in solving this type of biomolecular problem.

The quality of the results should be examined with respect to the objective, which is to achieve a negligible RMS compared to experimental ones, where the basic assumption is that the lower the energy the better the RMS. RMS, like any other tool has its own limitations. The user should consider which atoms should be superimposed. In table XIII RMS values are compared, where different atoms are selected for the superimposition. Calculating loops RMS between the N, Cα and C yielded an average RMS value of 1.86. When the carbonyl oxygen was added, the average RMS value was elevated to 2.10. Adding the protein's residues that are bonded to the loop increased the average RMS value to 2.62. One may assume that the inclusion of these two residues will reduce the RMS value because their coordinates are "correct". However, the opposite phenomenon is observed, at least in our test cases. In other words, when one superimposes the loops' atoms the RMS ignores the rest of the protein, and geometric factors, such as bond lengths and dihedral angles between the loop and the protein are ignored. Low RMS value does not necessarily indicate that the internal loop geometry is acceptable and one should verify that the predicted loops assume a reasonable geometry (the loop does not remain open), and do not clash with "known" parts of the protein. The phenomenon can be explained by the RMS superimposition mechanism. The RMS function translates and rotates the predicted loop to overlay the known loop, while

ignoring the rest of the protein. If one imagines a protein structure made of wire, a "wire loop" may be bent (by prediction) without changing the internal loop coordinates. Thus, for residues m to n one could achieve a RMS=0.0. On the other hand, this bend may cause a large deviation from the protein structure, so that attaching the "correctly predicted loop" into the protein will increase the RMS considerably, due to the deviation of the other residues from their protein positions.

Section 4: Examples of Other Biological Problems

The preceding sections gave in-depth test results for illustrating the efficacy of the present invention for solving a number of difficult biological problems. This section describes a number of other such problems, and how they could be solved with the present invention.

Homology Modeling. Homology modeling construction of unknown protein structures on the basis of proteins known from X-rays or from NMR studies requires "insertions" and "deletions" of peptide fragments as well as mutations compared to the known structure. The homologous parts of the target (to be constructed) are superimposed, residue by residue, over those of the known protein. Other parts may differ in length and are regularly encountered in loops, beta-hairpins and random coil parts of the known protein. Each such operation requires a re-evaluation of the backbone coordinates in those non-homologous parts, due to length differences ("insertions" and "deletions") as well as side chain positions, at least in the vicinity of the moderated part of the structure. Any planning of mutations in known protein structures may be aided by constructing models with an initial intact rigid backbone. Substantial progress in solving this acute problem has been already achieved by the method of the present invention.

The effects of "insertions" and "deletions" may be dealt with in the present invention by applying the above-described approach to the construction of loops, with or without concurrent positioning of side chains.

A few more issues, such as the employment of various force fields to evaluate the energy, as well as the use of alternative rotamer libraries, with and without statistical weights, may also be incorporated into the present invention for this problem.

Ring closure of peptides, peptidomimetics, and other cyclic structures. Many studies had indicated the potential therapeutic importance of cyclic ("conformationally restricted") peptides in preference to the biologically unstable linear peptides (Hruby, Life Sci. 1982; 31, 189; Altstein et al. J. Biol. Chem. 1999; 274:17573). Theoretical studies of the conformations of such peptides (Keasar & Rosenfeld, Folding and Design 1998; 3:379; Tieleman et al., Biophys. J. 1999; 76:1757) depend to a large degree on "correct" ring closure options due to high barriers between their conformations. Many such cyclic peptides have been studied in solution by NMR (Baysal & Meirovitch, Biopolymers 1999; 50:329).

Cyclization of active peptides and other linear molecules is one of the methods of choice for increasing their binding to biological receptors, due to the expected reduction in entropy loss, increasing their stability to digestion as well as strengthening their specificity and selectivity, etc.. The design of such cyclic structures may be aided considerably by preliminary modeling of the alternatives for ring closure. This is a function of many variables such as ring size, bond lengths, bond angles, and other factors.

This problem is quite similar to that of loop structure prediction with regard to the present invention. In general, cyclic peptides are smaller than loops, and so less "freedom" may be introduced into the conformational flexibility of the backbone and of side chains. Also, relatively small increments for phi and psi (backbone) angles are required for a thorough search for ring closure options.

Flexible docking of drug candidate to active sites. Computational methods for predicting the binding of ligands to their targets are generally based on seeking the most stable bound conformation of the complex (Stryndaka et al., Nature Struct. Biol. 1996; 3:233). Ideally, it will match the bound conformation that is observed crystallographically (Rosenfeld et al., Annu. Rev. Biophys. Biomol. Struct. 1995; 24:677; Clark & Westhead, Comput. Aided Mol. Des. 1996; 10:337; Abagyan & Totrov, J. Mol. Biol. 1994; 235:983; Trosset & Scheraga, Proc. Nat. Acad. Sci. USA 1998; 95:8011). However, low-energy conformations other than the global energy minimum of the complex could contribute to the binding affinity. Novel docking algorithms incorporate this assumption (Head et al., J. Phys. Chem. 1997; 101:1609) However, most of the flexible docking software such as DOCK, AUTODOCK, FLEXX, GOLD and others are not considering the flexibility of the target protein or biomolecule and do not consider the potential variations in active site protonation (pKa) state or in water content.

Flexible docking is essential for testing the ability of most molecules to bind to their biomolecular targets in different positions and variable modes of binding. In the interaction between a flexible drug and a protein, structural changes, mainly conformations, may occur in both the drug and the site of interaction in a protein (active sites of enzymes, binding site of a receptor protein).

With regard to the present invention, this is an extension of the problem of side chain positioning and also of determining a structure of a loop of a protein, differing from it in the need to move the drug by six degrees of freedom (translational + rotational) with respect to the biomolecular active site. The present invention must handle both the location of the side chains and the loops (backbone variations) predictions described earlier, but with the optimization applied to both a biomolecular target and a ligand at once, with the additional need to optimize their relative positions.

Those additional degrees of freedom may optionally be introduced as variables, but with special requirements. In addition, optionally and more preferably, the problem is analyzed according to the method of the present invention with the addition of an additional variable, which is the relative distance of the entities (the drug and the biomolecular active site, for example).

The variables thus include variables for distance and angles, for a total of six additional variables for translations and rotations. The present invention must handle both the location of the side chains and the loops (backbone variations) predictions described earlier, but with the optimization applied to both a biomolecular target and a ligand at once, with the additional need to optimize their relative positions. The variables thus preferably include variables for distance and angles, for a total of six such variables: three translations along X,Y,Z coordinate axes and three rotations about the same angles.

Structural comparisons of flexible molecules. The traditional RMS approach or other superimposition methods (Lemmen et al., Pac. Symp. Biocomput. 1999; 482) are inadequate for comparing a very large range of conformations for flexible molecules.

Such comparisons enable the assessment of the possibility that different molecules may be attached to the same biomolecular site/target. Two different molecules may display similar binding affinities to enzyme active sites or to a receptor. The method of the present invention enables the structural differences between such molecules to be optimized, in order to find candidates for a "bioactive conformation" of both.

This problem presents another conformational search for the present invention, but the function or quantitative parameter to be minimized in this case would be the RMS difference between spatial positions of selected atoms in the two molecules.

Construction of molecules from fragments. This is a classical problem in structure based drug design (Krygeer et al., Structure 1999; 7:297), that received much attention (Mizutani et al., J. Mol. Biol. 1994; 243:310; Tomioka & Itai, J. Comput. Aided Mol. Des. 1994; 8:347; Leach & Lewis, J. Comp. Chem. 1994; 15:233). Some excellent approaches to study the affinities of molecular fragments to biomolecules have been developed, such as GRID (Wade et al., J. Med. Chem. 1993; 36:140; Wade & Goodford, J. Med. Chem. 1993; 36:148; Boobbyer et al., J. Med. Chem. 1989; 32:1083), but the combination of fragments into molecules that may become drug candidates requires a vast computational effort. Again in this case, the pursuit should be after an ensemble of ligands and not a single one.

This is a major problem in drug design, where several positions of molecular fragments are known from other studies, while combining them into a specific and selective ligand is a complex task. Quite a few programs (such as GRID) (Wade & Goodford, J Med. Chem. 1993; 36: 148-156) are able to indicate the best interacting positions of a molecular fragment (such as a hydroxyl, an amine, a carbonyl, etc.) with an active site of a known protein's structure. However, the number of potential combinations of such fragments, in order to construct drug candidates or lead compounds, is very large and requires a process of optimization. In this case, the process must also be guided by chemical knowledge, in order to achieve structures that may be synthesized as well as being limited by molecular weight, lipophilic character etc.

This problem is different than all the others since chemical knowledge of synthesis and of molecular stability must be introduced into the evaluation of structures. The variables in this case would be the spatial positions and directions of fragments whose positions in an "active site" have been optimized previously, as well as molecular "connecting" fragments (such as aliphatic, alicyclic and aromatic) that would be employed to assemble the fragments into full viable structures and evaluate their energy of interaction with the "active site" as well as their internal energy and energy of hydration.

Small protein folding. In a recent short review on the "energy landscape in non-biological and biological molecules" (Fraunfelder & Leeson, Nature Struct. Biol. 1998;

5:757) the authors conclude that "Proteins that fold have been selected by evolution so that their energy landscapes resemble funnels". Those funnels have, in many cases, (Wales et al., Nature 1998; 394:758) a steep shape with low barriers inside, but other funnel shapes exist. For those proteins that can fold without chaperonins (Horovitz, Curr. Opin. Struct. Biol. 1998; 8:93) there may be many accessible conformations that are close to the "global minimum", which may be important for studying the dynamics and function. Searching for those funnels (Dill & Chan, Nature Struct. Biol. 1997; 4:10) became one of the central issues of modern biology, the "protein folding" problem. Following many years of studying small models with, mostly, on-lattice simulations (Shaknovitch, Curr. Opin. Struct. Biol. 1997; 7:29), more modern simulations attempt to fold peptide fragments or entire proteins (Dobson & Karplus, Curr. Opin. Struct. Biol. 1999; 9:92). Very long simulations (1μs) have been recently developed (Duan & Kollman, Science 1998; 282:740) and enabled the folding of a 36-residue protein. Monte Carlo methods (Hansmann & Okamoto, Curr. Opin. Struct. Biol. 1999; 9:177) and other stochastic dynamics (Sanderowitz & Still, J. Comput. Chem. 1998; 19:1294) are still popular. These energy based methods do not find easily the native fold of a protein with more than 35-40 residues.

Protein folding has been a central problem of biophysics in the last two decades. The method of the present invention may be applied to a set of proteins which have a relatively small number of residues, in the range of 50-80, depending on their primary structure. In addition to a "global" minimum, this approach can produce many other low energy conformations that are in the energy vicinity of the global minimum and contribute to the total character of the protein.

In a small protein of about 50 residues, the variables will be the phi and psi angles along the backbone (each with 6 or 12 rotations of 60° or 30° difference, respectively, as well as rotamers for the side chains. For the backbone alone, with 6 rotations for each phi and psi angle, the size of the problem is 6^{100} or $\sim 10^{66}$. With the additional rotamers that should be positioned simultaneously, it increases to about 10^{100}. Thus, although the resultant calculations may be complex, they can be performed with the method of the present invention.

It will be appreciated that the above descriptions are intended only to serve as examples, and that many other embodiments are possible within the spirit and the scope of the present invention.

Table I. 5PTI using "combined ensemble-stochastic approach"

	serial number of ensemble	number of segments	number of possible locations	number of possible combinations	value of combination with lowest energy in Kcal/mole	value of combination with highest energy in Kcal/mole
	1	2	7	10	31.1	33.5
	2	4	19	162	7.5	19.5
	3	3	20	280	12.7	28.7
	4	4	11	54	12.4	18.2
	5	2	17	66	-39.1	-14.7
	6	1	1	1	1.7	1.7
	7	1	3	3	3.9	4.6
	8	4	29	2,310	-40.7	-1.1
	*9	8	62	6,403,320	-263.2	2.9E+16
	10	1	2	2	15.6	15.6
	11	1	7	7	20.4	25.3
	12	3	30	968	-14.5	6.4
	13	2	8	15	23.8	29.5
	14	1	9	9	15.4	20.0
	15	1	2	2	-0.8	-0.2
	16	1	4	4	6.0	7.1
	17	1	8	8	12.8	15.9
	18	1	5	5	21.4	22.7
	19	2	8	15	13.8	19.5
	20	1	2	2	16.9	16.9
	21	1	2	2	22.0	22.2
total	21	45	256	6,407,245	-121.0	2.9E+16

* In ensemble number 9, there are 9 positions available for segment 1, 11 for segment 2, 4 for segment 3, 7 for segment 4, 7 for segment 5, 3 for segment 6, 10 for segment 7, and 11 for segment 8.

Table II. 5RSA using "combined ensemble-stochastic approach"

	serial number of ensemble	number of segments	number of possible locations	number of possible combinations	value of combination with lowest energy in Kcal/mole	value of combination with highest energy in Kcal/mole
	1	1	1	1	20.6	20.6
	2	8	54	1,626,240	-139.6	-117.7
	3	1	1	1	29.9	29.9
	4	4	25	1,250	-10.5	-1.2
	5	3	8	16	4.7	16.1
	6	3	9	27	34.6	36.8
	7	6	36	16,632	-14.6	7.3
	8	2	7	12	31.4	32.8
	9	1	1	1	42.4	42.4
	10	7	50	611,520	-69.2	-36.3
	11	2	12	36	-24.9	-17.2
	12	3	17	144	-10.8	-5.3
	13	2	9	14	2.5	9.4
	14	4	16	180	-23.5	-15.8
	15	1	3	3	8.2	10.5
	16	1	2	2	6.6	7.3
	17	1	1	1	7.8	7.8
	18	4	26	1,440	-37.7	-21.0
	19	1	3	3	10.3	12.5
	20	3	15	80	2.1	29.0
	21	2	6	5	44.9	49.8
	22	1	1	1	6.7	6.7
	23	1	3	3	8.1	10.0
	24	3	17	510	38.3	45.6
	25	3	28	120	33.7	41.9
	26	1	6	6	12.8	14.8
	27	1	2	2	19.8	20.0
	28	1	2	2	24.0	24.2
	29	5	37	17,248	-122.6	-91.4
	30	2	13	42	-19.7	-7.5
	31	4	17	300	-91.6	-59.6
	32	2	14	49	-8.6	-1.3
	33	1	2	2	7.3	7.3
	34	2	11	24	24.9	49.0
	35	1	3	3	18.8	19.5
	36	1	4	4	14.2	16.2
	37	3	23	448	57.9	68.1
total	37	92	485	2,276,372	-60.8	261.3

Table III. 2MB5 using "combined ensemble-stochastic approach"

	serial number of ensemble	number of segments	number of possible locations	number of possible combinations	value of combination with lowest energy in Kcal/mole	value of combination with highest energy in Kcal/mole
	1	3	11	25	136.5	142.9
	2	2	5	6	61.9	64.9
	3	1	2	2	57.0	58.5
	4	3	26	648	132.4	148.6
	5	1	3	3	22.9	26.1
	6	5	20	720	136.2	154.1
	7	2	7	12	78.1	83.5
	8	1	4	4	16.9	18.9
	9	3	20	216	96.3	107.4
	10	4	27	1,764	104.3	116.1
	11	7	39	25,872	273.9	294.2
	12	2	5	5	77.8	81.4
	13	1	2	2	57.8	58.4
	14	1	1	1	54.8	54.8
	15	4	10	16	66.4	70.2
	16	1	4	4	31.7	35.6
	17	1	1	1	25.2	25.2
	18	1	3	3	66.1	66.2
	19	1	1	1	46.3	46.3
	20	1	3	3	27.1	29.6
	21	1	1	1	28.5	28.5
	22	3	7	12	119.6	122.2
	23	1	3	3	21.7	24.8
	24	3	19	200	83.0	90.4
	25	2	9	20	65.6	68.8
	26	3	16	140	63.9	78.5
	27	1	7	7	37.7	38.9
	28	1	7	7	67.8	72.9
	29	1	7	7	34.2	37.7
	30	1	6	6	52.7	57.8
	31	1	3	3	49.9	52.4
	32	5	44	48,510	115.9	137.3
	33	1	7	7	59.3	63.6
	34	7	51	362,880	194.4	222.2
	35	1	4	4	44.1	45.4
	36	1	3	3	70.4	70.4
	37	2	10	24	63.2	74.4
	38	1	3	3	45.6	45.7
	39	1	2	2	61.7	61.7
	40	2	19	88	43.9	47.8
	41	1	4	4	65.7	66.1
	42	1	7	7	35.0	38.7
	43	1	5	5	35.5	40.0
total	43	87	438	441,251	3,029.0	3,269.1

Table IV. 1NTP using "combined ensemble-stochastic approach"

	serial number of ensemble	number of segments	number of possible locations	number of possible combinations	value of combination with lowest energy in Kcal/mole	value of combination with highest energy in Kcal/mole
	1	1	4	4	4.8	6.0
	2	1	2	2	6.8	8.0
	3	1	4	4	17.3	22.8
	4	2	4	4	35.0	37.8
	5	1	1	1	7.7	7.7
	6	1	3	3	12.4	15.0
	7	2	4	4	18.9	21.3
	8	1	1	1	13.8	13.8
	9	1	1	3	16.6	18.1
	10	1	1	1	24.0	24.0
	11	1	1	1	14.3	14.3
	12	1	4	4	4.9	7.4
	13	1	1	1	15.5	15.5
	14	1	1	1	6.5	6.5
	15	1	1	1	18.0	18.0
	16	1	2	2	9.0	11.1
	17	1	1	1	10.0	10.0
	18	2	4	3	41.0	45.5
	19	1	1	1	9.9	9.9
	20	1	1	1	37.8	37.8
	21	1	6	6	11.3	13.5
	22	1	2	2	10.0	10.6
	23	1	3	3	6.4	10.4
	24	1	4	4	11.2	12.6
	25	1	3	3	15.2	15.4
	26	1	3	3	11.0	12.0
	27	2	5	4	11.6	12.5
	28	2	6	8	11.5	16.2
	29	1	1	1	14.3	14.3
	30	1	1	1	28.3	28.3
	31	1	7	7	10.8	12.9
	32	1	3	3	9.0	9.8
	33	1	1	1	9.1	9.1
total	33	38	87	89	483.9	528.2

Table V. 1IXH using "combined ensemble-stochastic approach"

serial number of ensemble	number of segments	number of possible locations	number of possible combinations	value of combination with lowest energy in Kcal/mole	value of combination with highest energy in Kcal/mole
1	1	5	5	-9.1	-6.0
2	2	14	40	-40.4	-25.0
3	3	13	80	-53.7	-42.4
4	1	1	1	-3.0	-3.0
5	1	4	4	-7.9	-7.1
6	1	4	4	-8.2	-5.3
7	1	3	3	-15.4	-12.2
8	2	7	10	-36.2	-21.3
9	1	3	3	-2.4	-0.4
10	1	2	2	-3.0	-2.7
11	1	2	2	-2.3	-2.2
12	1	1	1	-21.7	-21.7
13	3	9	20	-27.2	-21.9
14	1	1	1	-19.5	-19.5
15	2	4	4	-33.3	-30.1
16	2	4	4	-4.4	-3.7
17	1	1	1	-1.1	-1.1
18	1	1	1	2.5	2.5
19	1	1	1	-4.4	-4.4
20	1	1	1	-4.9	-4.9
21	4	19	378	-67.6	-38.8
22	1	4	4	-6.2	-4.6
23	1	2	2	-6.5	-6.1
24	1	5	6	-49.9	-48.5
25	1	1	1	-0.4	-0.4
26	1	2	2	-47.9	-47.8
27	1	2	2	-5.6	-5.4
28	1	3	3	-5.2	-3.9
29	1	1	1	-18.8	-18.8
30	1	6	6	-14.1	-12.0
31	1	2	2	-34.2	-33.8
32	1	2	2	-8.9	-8.5
33	1	1	1	-5.3	-5.3
34	1	4	4	-11.5	-10.5
35	1	2	2	-13.8	-13.0
36	1	5	5	-11.1	-8.5
37	1	1	1	-7.1	-7.1
38	1	1	1	-18.2	-18.2
39	1	1	1	-12.7	-12.7
40	1	2	2	-11.1	-11.0
41	1	2	2	-33.6	-33.5
42	1	1	1	-19.2	-19.2
43	1	3	3	-9.6	-9.2
44	2	4	4	-1.3	-0.3
45	2	4	4	-3.9	-2.2
total	45	58	161	628	-719.2	-611.7

Table VI. RMS values summary

Protein	Method	RMS differences ser	tyr	thr	water
5RSA	number of hydrogens	15	6	10	180
	CVFF minimization	0.98	0.88	1.21	1.19
	Brunger and Karplus	0.98	0.60	1.12	1.2[&&]
	Bass et al.	0.61	0.60	0.30	*
	Combined ensemble-stochastic	0.60	0.39	0.43	0.64
	Pure stochastic	0.56	0.48	0.56	0.65
5PTI	number of hydrogens	1	4	3	108
	CVFF minimization	1.08	1.03	1.19	1.26
	Brunger and Karplus	0.71	0.81	0.19	0.35**
	Bass et al.	0.96	*	0.07	*
	Combined ensemble-stochastic	0.40	0.72	0.41	0.69
	Pure stochastic	0.30	0.58	0.41	0.67
2MB5	number of hydrogens	6.00	3.00	5.00	178.00
	CVFF minimization	0.64	0.73	0.96	1.22
	Brunger and Karplus	*	*	*	*
	Bass et al.	*	*	*	*
	Combined ensemble-stochastic	0.42	0.36	0.40	0.64
	Pure stochastic	0.44	0.40	0.38	0.68
1NTP	number of hydrogens	33	10	10	#
	CVFF minimization	1.04	0.70	0.51	#
	Brunger and Karplus	0.89	0.64	0.34	#
	Bass et al.	0.34	0.44	0.17	#
	Combined ensemble-stochastic	0.64	0.46	0.36	#
	Pure stochastic	0.53	0.45	0.27	#
1IXH	number of hydrogens	19	12	21	#
	CVFF minimization	0.63	0.74	1.07	#
	Brunger and Karplus	*	*	*	#
	Bass et al.	*	*	*	#
	Combined ensemble-stochastic	0.43	0.44	0.57	#
	Pure stochastic	0.50	0.48	0.57	#

* RMS values were not reported.
** Calculation included only 4 water molecules (8 hydrogens).
# There are no water hydrogens in the neutron diffraction structure.
&& Calculation included 128 water molecules (256 hydrogens).

Table VII. RMS[a] Values for the stochastic search

Name	PDB code	size	Combinations[b]	RMS of global energy conformation	Average RMS[c] for 1000 lowest lowest energy conformers	Best possible RMS for this library	Average energy gap from global minimum for 1000 lowest energy conformers[c,d]
1. Crambin	1crn	46	$1.04*10^{14}$	1.32	1.38 (1.27-1.59)	0.99	1.93 (0-2.49)
2. Ribosomal protein	1ctf	68	$2.48*10^{34}$	2.31	2.33 (2.27-2.42)	0.94	3.28 (0-4.96)
3. Complement control protein	1hcc	59	$5.36*10^{26}$	2.17	2.18 (2.04-2.26)	1.20	1.25 (0-1.65)
4. Ovomucoid third domain	2ovo	56	$3.42*10^{24}$	1.94	2.03 (1.89-2.22)	1.24	3.64 (0-5.52)
5. Erabutoxin B	3ebx	62	$2.39*10^{31}$	2.48	2.50 (2.42-2.56)	1.20	1.35 (0-1.93)
6. Rubredoxin	5rxn	54	$1.95*10^{27}$	2.17	2.20 (2.16-2.25)	1.52	1.64 (0-2.24)
7. Lysozyme	2ihl	129	$1.70*10^{62}$	2.27	2.26 (2.22-2.30)	1.27	1.98 (0-2.56)
8. Ribosomal protein	1whi	122	$2.80*10^{73}$	2.50	2.48 (2.40-2.55)	0.95	3.33 (0-4.44)
9. Endonuclease	2end	137	$1.15*10^{82}$	2.60	2.68 (2.66-2.76)	1.41	3.03 (0-3.96)
10. Hydrolase	1arb	263	$2.29*10^{105}$	2.22	2.24 (2.21-2.28)	1.30	0.55 (0-0.74)
Average				2.20	2.23	1.20	2.20

[a]RMS values for all non hydrogen side chain atoms excluding C
[b]number of conformations after backbone clashes are relieved.
[c]values in brackets indicate the minimal and maximal values.
[d]values given in Kcal/mole

Table VIII. Residues adopting multiple conformations in E. coli ribonuclease HI.

Type	Angle	1rnh	2rn2	NMR	MD	Our results	%	Score
M1			160	-174	178	-177	33.33%	+
				-76	-60	-62, -66	66.67%	+
M1			-75	-80	-79	-70	33.00%	+
				-159	-164	170	33.00%	?
				163	168	170	33.00%	+
						78	33.00%	-
L2			-175	163	-173	176	50.00%	+
				-66	-65		0.00%	(*)
		56				78	50.00%	(1rnh)
F8		-59	-60	-69	-56	-63	100.00%	(+)
F8		-89	-95	-67	-90		0.00%	(*)
				93		98	100.00%	(NMR)
L14			164	167		-177	50.00%	+
				72	62	78	50.00%	+
Y22		61	77	66	67	61	100.00%	(+)
Y22		80	67	106	88	90	100.00%	(+)
I25		-67	-64	-61	-67		0.00%	*
						-175	50.00%	-
						59	50.00%	-
R29			-64	-56	-60	-66	50.00%	+
					-173	-178	50.00%	(MD)
R29		-70	-97	-59	-66	-66, -72	28.57%	+
				175	-176	-175, -179, -177, 179	57.14%	+
					76	67	14.29%	(MD)
R29				-172	-175	-172	14.29%	+
			-94	-75	-93	-87, -90	28.57%	+
				84	103	89	14.29%	+
		134			134	167, 172	42.86%	?
R31			-71	-74	-63	-63	100.00%	+
		-170		-173	-169		0.00%	(*)
K33		-99	-101	-72	-66	-67	50.00%	+
				-170	-174	178	50.00%	+
K33			-154	-176	179	-178	33.33%	+
				67	63	63	33.33%	+
		-37			-69	-70	33.33%	+
R41		165	-179	-178	-172	-179, 179, 177	75.00%	+
				-66	-72	-66	25.00%	+
R41				-176	-172	-177	25.00%	+
		88				86, 89	50.00%	(1rnh)
			-101	-75	-107	-86	25.00%	+
R46		-62	-70	-78	-70	-72	100.00%	(+)
R46		-174	-174	180	177	-179	50.00%	+
				-97		-72	50.00%	(NMR)
M47		-77	-72	-80	-65	-70	100.00%	(+)
M47		162	88	104	97	99	100.00%	(+)
L49			-92	-100		-73	100.00%	+
		-158		158	-176		0.00%	*
L49			-177	173	178	176	50.00%	+
		45				43	50.00%	+
M50			70	64	62		0.00%	*
		180		167	170	179	100.00%	+
M50			76	57	89	72	100.00%	(+)
L56		-69	-65	-67	-70	-74	100.00%	+

				75	-170		0.00%	(*)
E57		-69	-64	-78	-73	-72	50.00%	+
				-166	-169	-174	50.00%	+
K60			-50	-79	-60	-77, -88	50.00%	+
		-175		-179	-175	-180, -169	50.00%	+
K60		132	175	178	174, -169	179, -172	50.00%	+
				-71		-64	25.00%	(NMR)
					70	76	25.00%	(MD)
E61		178		-173	-178	179	30.00%	+
			-101			-70	30.00%	+
				63	76	82	30.00%	+
Q72		162	-177	-165	-172		0.00%	*
				-69	-62		0.00%	*
					67	63	100.00%	(MD)
Y73		61	75	65	70	98	100.00%	(+)
K87				-175	-175	-170, -171	75.00%	+
						65	25.00%	-
		-69	-93	-70	-69		0.00%	*
K87			-93	-73	-68		0.00%	*
		-175		-175	180	180, -178, 177	75.00%	+
					74	66	25.00%	(MD)
K91		167	168	-179	180		0.00%	-
				-72		-70, -72	100.00%	(NMR)
					69		0.00%	(MD)
K95				175	-176	-179, -174, 179	75.00%	+
						72	25.00%	-
		-69	-53	-80	-67		0.00%	*
K95		155	178	178	180	179, 180	66.67%	+
				75	65		0.00%	*
				-74	-65	-70	33.33%	+
K99			-166	174	177	180, -174	50.00%	+
		58		77	67	72	25.00%	+
					-85	-68	25.00%	(MD)
K99		132	-24	173	179	-179	100.00%	+
				74	66		0.00%	*
				-78	-69		0.00%	*
K99		121	87	69	65		0.00%	*
				-73	-63	-70, -72	50.00%	+
				-175	-179	178, 179	50.00%	+
N100		-55	-67	-54	-60	-65, -67	66.00%	+
				-156		-168	33.00%	(NMR)
N100			-32	-28	-46	-64	50.00%	+
						-109	50.00%	-
W104		-78	-76	-81	-74	-61	100.00%	(+)
W104		118	119	124	115	98	100.00%	(+)
Q105		-60	-74	-78	-73		0.00%	*
				-156		-170	100.00%	(NMR)
Q105		-172	-168	174	-177	180	100.00%	(+)
R106			-178	-153	-171	-175	66.67%	+
		-90		-76		-73	33.33%	+
H114		-58	-54	-97	-64	-72	100.00%	(+)
H114		-72	-77		-61		0.00%	(*)
				-30		-13	50.00%	(NMR)
				83, 122		99	50.00%	(NMR)
Q115		-44	-61	-64	-64		0.00%	*
				-170	-172	-173	100.00%	+

K117		144	149	180	178	-172	25.00%	+
				75	64	76	25.00%	+
						-67, -74	50.00%	-
K122			-82	-64	-63	-65	50.00%	+
				-162	-169	-177	50.00%	+
		31			62		0.00%	(*)
K122				171	178	-178	100.00%	+
		-92	-85	-76	-67		0.00%	*
K122		178	-167	-170	-179	-178	50.00%	+
					-69	-69	50.00%	+
K122			-178	178	-179	-178	50.00%	+
				73	68		0.00%	*
		-59			-81	-67	50.00%	+
H124			60	65	53, 73		0.00%	(*)
				172		-173	100.00%	(NMR)
		-85			-64		0.00%	*
C133		-76	-75	-60	-66	-64	100.00%	(+)
E135		-49		-80	-65	-69	50.00%	+
			-159	-140	-170	-175	50.00%	+
R138				-167	-160	-175	50.00%	+
		-61	-67	-71	-73	-72	50.00%	+
M142		-64	-74	-65	-71	-72	50.00%	+
				-150	-171	-177	50.00%	+
M142		180			178	179	25.00%	+
				78	58	67	25.00%	+
			-59	-72		-66, -68	50.00%	+
M142			-38	-65	-71		0.00%	*
		164		167	180	-175, 170	50.00%	+
				88	70	73, 99	50.00%	+
N143		-163	176	-169	-167		0.00%	*
				-75	-62	-74	100.00%	+
E154		-	-62	-68	-63	-72	50.00%	+
					-171	-173	50.00%	(MD)

* angle not found by the algorithm due to force field limitation
(*) angle was not included in the search (missing in rotamer library)
(1mh) angle reported in 1inh crystal structure
(NMR) angle reported in NMR model
- calculated angle was found neither in crystal structures nor in NMR or MD
+ accurate result
(+) accurate result (see text)
? angle deviates in ~40°, thus result is ambiguous

Table IX. Results of the stochastic algorithm versus the DEE/A* algorithm

Search strategy	The test case	atom model	rotamer library	RMS of global energy conformation[a]	Best possible RMS for this library	Maximal number of residues	Maximal number of combinations[b]
Stochastic algorithm	same proteins DEE/A* reported results normal electrostatics	All atom model	SCWRL library	2.07 (1.32-2.48)	1.18	68	$2.48*10^{34}$
	Large proteins normal electrostatics	All atom model	SCWRL library	2.40 (2.22-2.60)	1.23	263	$2.29*10^{105}$
DEE/A* algorithm[c]	normal electrostatics	All atom model	Lavery library	1.92 (1.74-2.33)	0.83	68	10^{43}
	normal electrostatics	United atom model	Lavery library	1.84 (1.48-2.11)	0.83	68	10^{43}
	reduced electrostatics	All atom model	Lavery library	1.97 (1.66-2.04)	0.83	68	10^{43}
	reduced electrostatics	United atom model	Lavery library	1.83 (1.48-2.11)	0.83	68	10^{43}
	normal electrostatics	All atom model	Desmet library	1.84 (1.24-2.16)	0.75	68	10^{43}
	normal electrostatics	United atom model	Desmet library	1.76 (1.18-2.16)	0.75	68	10^{43}
	reduced electrostatics	All atom model	Desmet library	1.72 (0.11-2.13)	0.75	68	10^{43}
	reduced electrostatics	United atom model	Desmet library	1.77 (1.26-2.26)	0.75	68	10^{43}

[a]values in brackets indicate minimal and maximal values.

[b]number of possible combinations after exclusion of rotamers that clash with the backbone.

[c]see ref. Of Leach & Lemon (1998)

TableX. bacteriorhodopsin loops

Loop number	size	remarks
1. met 32-ser 35	4 residues	Predicted by our algorithm
2. tyr 64-tyr 79	16 residues	Predicted by our algorithm
3. asp 102-ala 103	2 residues	Can be solved by molecular modeling tools
4. lys 129	1 residue	Can be solved by molecular modeling tools
5. phe 156-met 163	8 residues	PDB entry does not contain the coordinates
6. ser 193-val 199	7 residues	Predicted by our algorithm

Table XI. Comparison to experimental results: bacteriorhodopsin loops

Pdb code	loop residues	length	sequence	rms backbone[1]
1c3w	32-35	4	met, gly, val, ser	0.28
1c3w	193-199	6	ser, glu, gly, ala, gly, ile, val	1.31
1c3w	64-79	16	tyr, gly, leu, thr, met, val, pro, phe gly, gly, glu, gln, asn, pro, ile, tyr	2.46
Average				1.35

[1]N, C alpha, and C rmsd values given for the top prediction of three methods

Table XII. Comparison to other methods

Pdb code	loop residues	length	sequence	rms backbone[4] van Vlijmen& Karplus method	Deane & Blundell	Glick & Goldblum
2apr	76-83	8	ser, tyr, gly, asp, gly, ser, ser, ala	5.2	2.6	2.85
8abp	203-208	6	gly, met, asn, asp, ser, thr	0.3	2.5	1.86
2act	139-144	6	ala, ala, gly, asp, ala, phe	1.6	1.5	2.99
3grs	83-89	7	ala, asp, tyr, gly, phe, pro, ser	4.6	2	1.06
5cpa	231-237	7	lys, ser, leu, tyr, gly, thr, ser	2.1	1.3	1.24
2fb4	H26-H32	7	gly, phe, ile, phe, ser, ser, tyr	1.6	1.9	1.69
2fbj	H100-H106	7	his, tyr, tyr, gly, tyr, asn, ala	0.5	3.2	1.36
Average				2.3	2.1	1.86

[1]N, C alpha, and C rmsd values given for the top prediction of three methods

Table XIII. Comparison of different RMS meausrments for the same loops

Pdb code	length	RMS[1]	RMS[2]	RMS[3]
2apr	8	2.85	2.89	3.24
8abp	6	1.86	2.24	2.67
2act	6	2.99	3.19	3.89
3grs	7	1.06	1.11	2.24
5cpa	7	1.24	1.41	1.36
2fb4	7	1.69	2.29	2.42
2fbj	7	1.36	1.6	2.54
Average		1.86	2.10	2.62

[1]RMS between N, C alpha, and C atoms of the predicted and experimetal loops

[2]RMS between N, C alpha, C and O atoms of the predicted and experimetal loops

[3]RMS between N, C alpha, C and O atoms of the predicted and experimetal loops including the n-1 and n+1 residues in the protein

WHAT IS CLAIMED IS:

1. A method for searching through combinatorial space, the space featuring multiple combinations, each combination being composed of a plurality of elements, the steps of the method being performed by a data processor, the method comprising the steps of:

(a) providing a quantitative parameter for determining success of a result of a search through the combinatorial space, said quantitative parameter being measurable for each combination;

(b) selecting a plurality of combinations in the combinatorial space to form selected combinations;

(c) calculating a value for said quantitative parameter for each of said plurality of selected combinations;

(d) determining an effect of each element on said value of said quantitative parameter; and

(e) retaining at least one combination according to said effect, to provide a result of searching through the combinatorial space.

2. The method of claim 1, further comprising a step being performed before step (a):

determining a structure for the multiple combinations of the combinatorial space, such that an interaction exists between the elements.

3. The method of claim 2, wherein each element is a variable having a value, and said quantitative parameter is calculated according to said values of said variables for each combination and said interaction between said variables.

4. The method of claim 3, wherein said quantitative parameter is a cost function.

5. The method of claim 4, wherein each variable has a single discrete value for any particular combination.

6. The method of claims 1 or 5, wherein step (e) further comprises the step of:

(i) discarding a value if said value does not consistently improve said cost function; and

(ii) discarding each combination featuring said value for said variable.

7. The method of claim 6, wherein step (e) further comprises the steps of:

(iii) determining if a number of remaining combinations is below a minimum number; and

(iv) if said number of remaining combinations is above said minimum number, repeating steps (c) - (e) at least once.

8. The method of claim 7, wherein step (e) further comprises the step of:

(v) if said number of remaining combinations is below said minimum number, evaluating each remaining contribution according to a parameter.

9. The method of claim 8, wherein step (v) is performed with an exhaustive search of said remaining combinations.

10. The method of claim 6, wherein step (i) is performed according to the steps of:

(1) creating a plurality of combinations by assigning values randomly to said variables;

(2) calculating said value for said cost function; and

(3) if said value is found in a plurality of combinations having a value for said cost function being below a predetermined minimum value, determining said effect of said value to be a negative effect.

11. The method of claim 10, wherein step (3) is performed wherein said value for said cost function is determined to be below said predetermined minimum value, if said value for said variable is found in combinations having said value for said cost function below said predetermined minimum value, and said value for said variable is not found in combinations having said value for said cost function above a predetermined desirable value.

12. The method of claim 6, wherein each value is for a location for a polar proton in a biological molecule, such that said cost function is a minimized energy calculation for said polar protons.

13. The method of claim 12, wherein the combinations are determined according to the steps of:

(i) parameterizing atoms of said biological molecule;

(ii) dividing hydrogen atoms and lone pairs into three categories: trivial hydrogen atoms; polar hydrogen atoms; and non-trivial lone pairs; and

(iii) adding trivial hydrogen atoms to each combination.

14. The method of claim 13, wherein said cost function is calculated according to a pairwise non-bonding energy function:

$$E(r_{i,j}) = \sum_{i<j} \left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} + \frac{q_i q_j}{\varepsilon * r_{i,j}}\right)$$

wherein $A_{i,j}$ is the repulsion parameter for the two (i, j) atoms, $B_{i,j}$ is their attractive polarizability parameter, q_i is the partial charge, $r_{i,j}$ is the distance between atoms, and ε is the dielectric constant.

15. The method of claim 6, wherein each combination of the combinatorial space includes locations for side chains of amino acids in a protein, such that said cost function is a minimized energy calculation for said side chains of amino acids.

16. The method of claim 15, wherein each combination is formed from rotamers for each side chain, such that the step of forming the combinations includes the step of eliminating rotamers clashing with a backbone of said protein.

17. The method of claim 15, wherein each element is a rotamer and said effect of said element on each combination is determined by sampling a plurality of combinations and determining a distribution of said quantitative parameter across said sampled plurality of combinations for each ensemble.

18. The method of claim 6, wherein each combination of the combinatorial space includes a structure for a loop in a protein, such that said cost function is a minimized energy calculation for said structure of said loop.

19. The method of claim 18, wherein said structure for said loop is determined according to a plurality of pairs of angles between residues of said protein.

20. The method of claim 19, wherein a value for each pair of angles is randomly selected to form each combination, each combination having an associated value for said cost function, and wherein said value is removed if said value contributes only to combinations having associated values above a predetermined threshold.

21. The method of claim 20, wherein step (e) further comprises the step of evaluating each combination according to an existence of a clash between side chains of amino acids in said loop, such that if said clash exists, said combination is discarded.

22. The method of claim 6, wherein each combination of the combinatorial space includes a structure for every loop in a target protein, such that said cost function is a minimized energy calculation for said structure of said loops, said structure for said loops being determined according to a plurality of pairs of angles between residues of said protein, wherein step (a) includes the step of providing a defined structure for a known protein, said known protein being homologous to said target protein, and wherein step (e) further comprises the step of evaluating each combination according to an existence of a clash between side chains of amino acids in said loops in said target protein and between said loops and a remainder of said target protein after said loops have been evaluated by comparison to said structure of said known protein, such that if said clash exists, said combination is discarded, such that said structure for said target protein is determined according to said structure for said known protein.

23. The method of claim 22, wherein a value for each pair of angles is randomly selected to form each combination, each combination having an associated value for said cost function, and wherein said value is removed if said value contributes only to combinations having associated values above a predetermined threshold.

24. The method of claim 6, wherein each combination of the combinatorial space includes locations for a plurality of moieties in a cyclized molecule, said structure of said cyclized molecule being determined from a structure of a linear molecule, such that said cost function is a minimized energy calculation for said locations of said moieties by comparison to said structure of said linear molecule.

25. The method of claim 6, wherein each combination of the combinatorial space includes an assembly of molecular fragments to form a single molecule, said assembly featuring a structure for linking each molecular fragment to at least one other molecular fragment at a defined location, such that said cost function is a minimized energy calculation for said defined locations of said molecular fragments in said structure of said assembly.

26. The method of claim 6, wherein each combination of the combinatorial space includes at least a portion of a structure of a first entity and of a second entity, each portion being defined according to variables for rotations about angles between each said portion and a relative distance between said first entity and said second entity, said relative distance being defined according to variables for translations along coordinate axes, such that said cost function is a minimized energy calculation for an interaction between said first entity and said second entity, for said distance, and for said at least a portion of said first entity and said second entity.

Figure 1

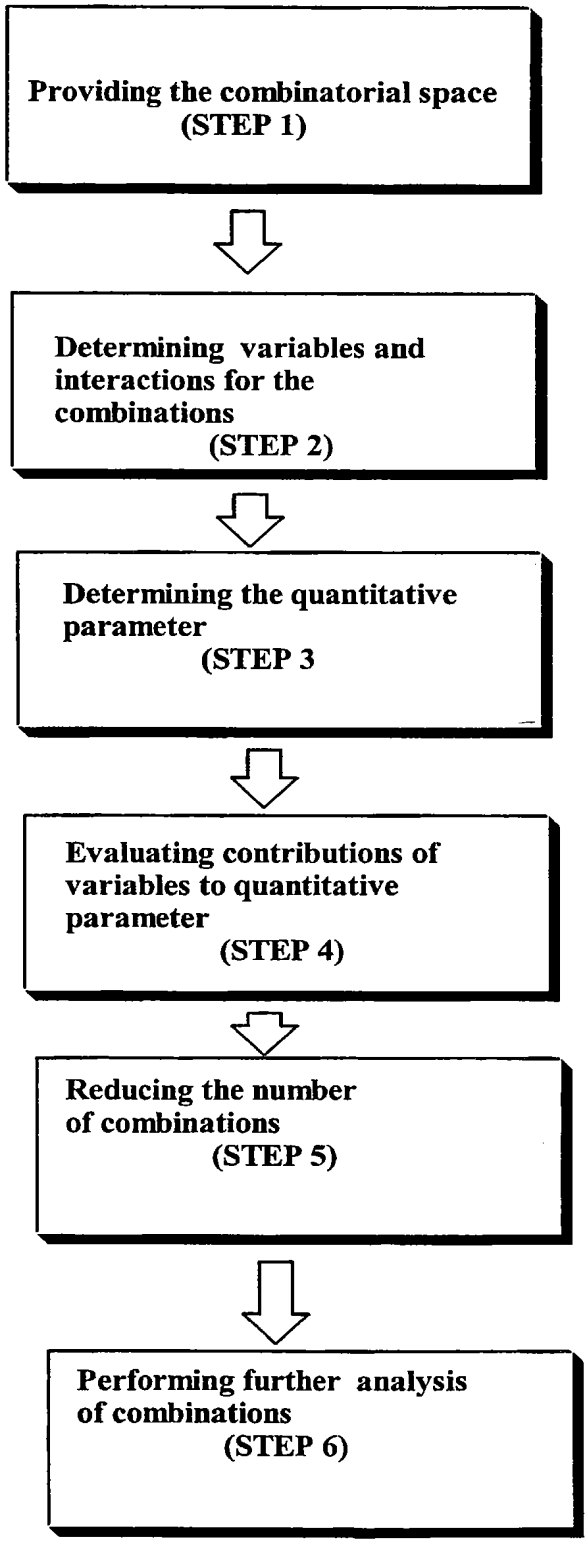


Figure 2

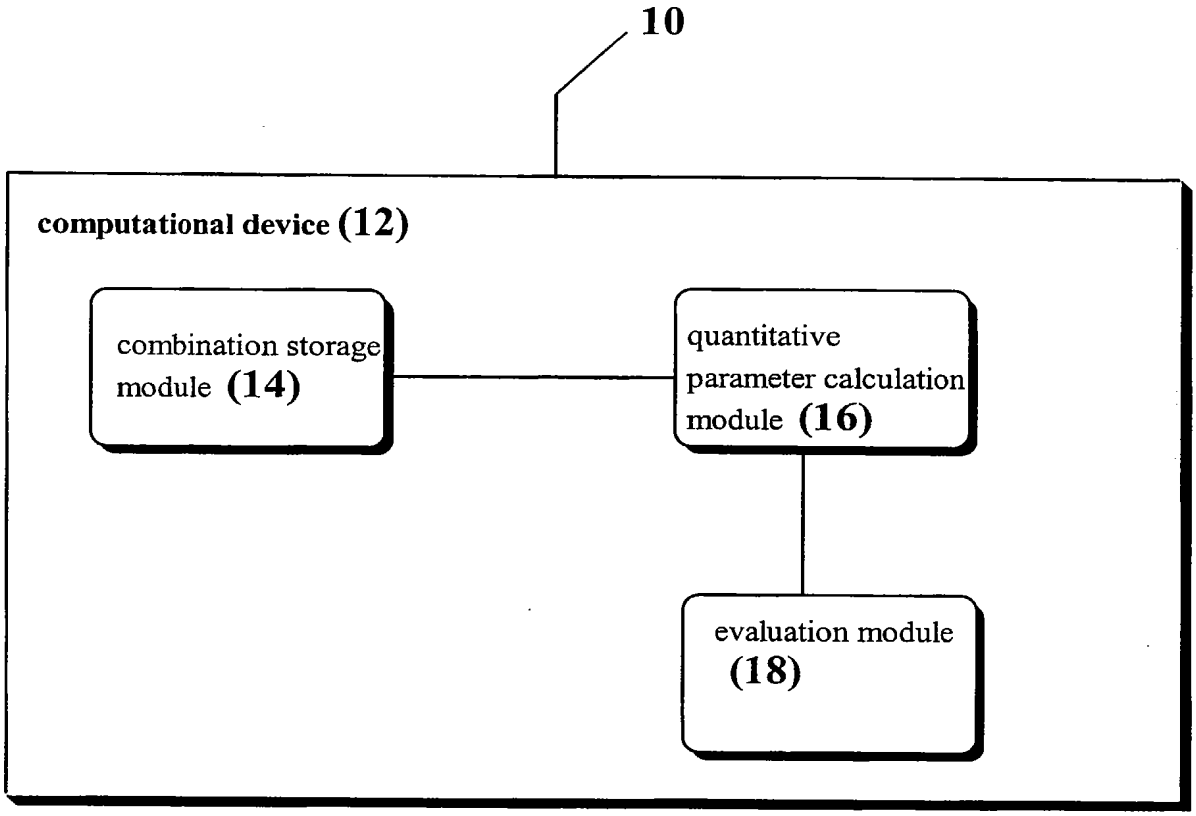


Figure 3

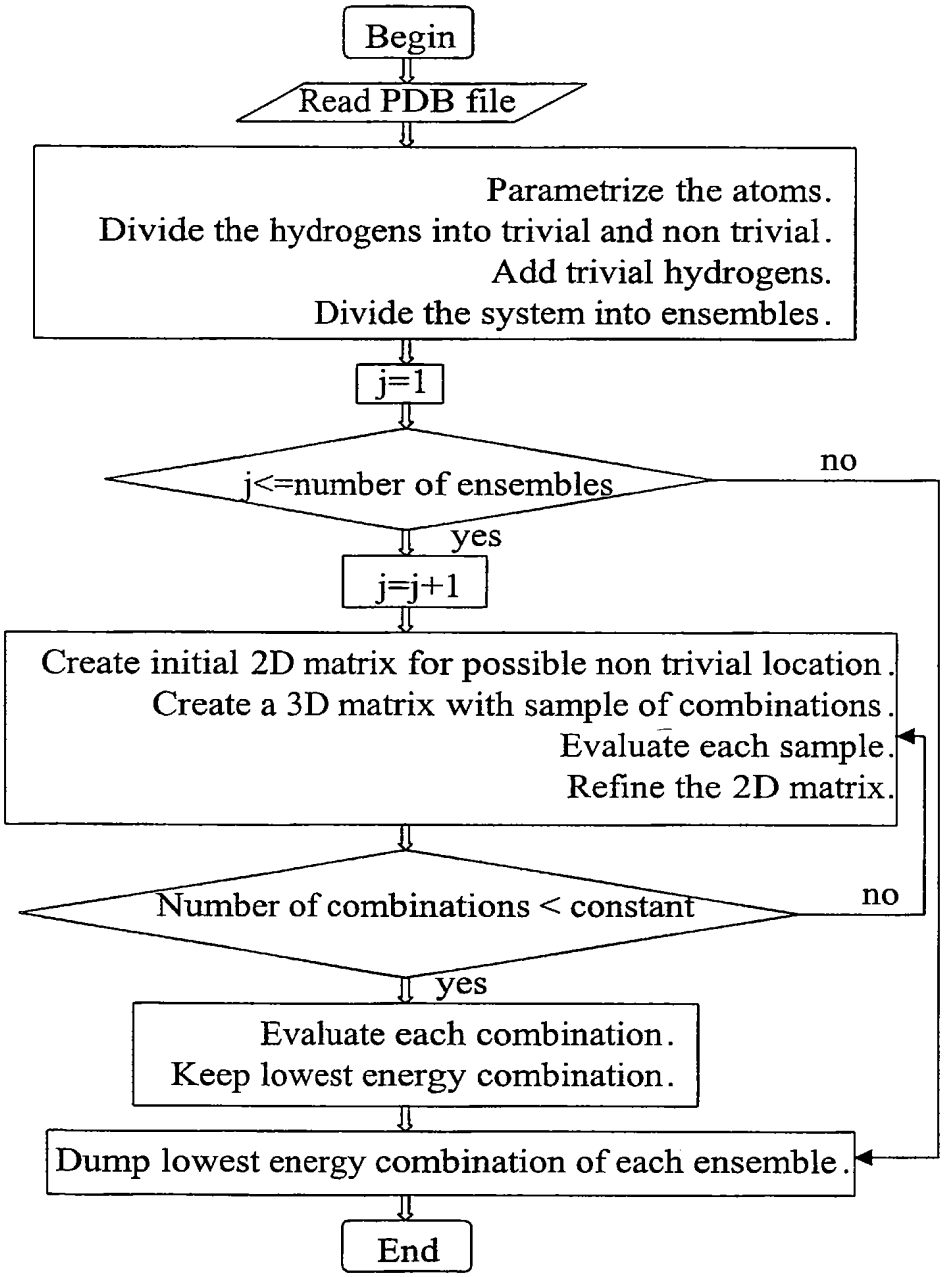


Figure 4

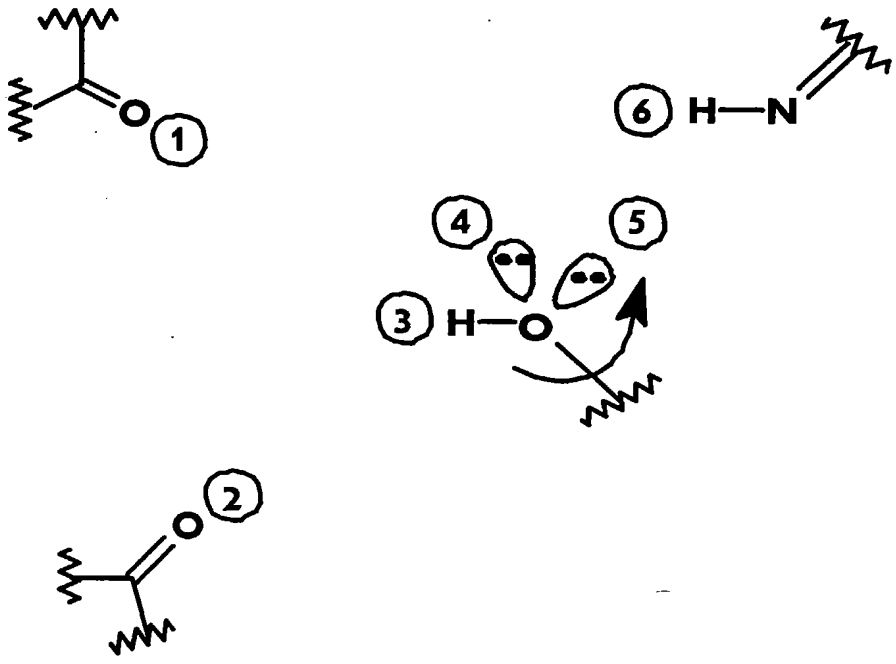


Figure 5

Figure 5A

rotatable atom number	3	3	4	5
trivial atom number	1	2	6	6

Figure 5B

rotatable atom number	3	3	4	
trivial atom number	1	2	6	

Figure 5C

rotatable atom number	3		4	
trivial atom number	1		6	
rotatable atom number		3	4	
trivial atom number		2	6	

Figure 6

The refined 2D matrix		segment 1	segment 2		segment d_0
	rotatable atom number	5 5 5 5 6	23 23 24 24	...	2334 2334 2335
	trivial atom number	1 2 3 4 8	34 55 59 1011	...	323 434 433
A. The n systems sampled					
sample number					
1st	rotatable atom number	5	23	...	2334
	trivial atom number	1	34	...	323
2nd	rotatable atom number	5	23	...	2334
	trivial atom number	1	34	...	434
3rd	rotatable atom number	6	24	...	2334
	trivial atom number	8	1011	...	434
.	.			...	
nth	rotatable atom number	6	24	...	2335
	trivial atom number	8	1011	...	433
B. The n highest energy systems (*H*)					
sample number					
1st	rotatable atom number	5	23	...	2334
	trivial atom number	1	34	...	323
2nd	rotatable atom number	5	23	...	2334
	trivial atom number	1	34	...	434
3rd	rotatable atom number	5	23	...	2334
	trivial atom number	1	34	...	434
nth	rotatable atom number	5	23	...	2334
	trivial atom number	1	34	...	434
C. The vector h					
	rotatable atom number	5	23	...	0
	trivial atom number	1	34	...	0
D. The vector l					
	rotatable atom number	5 5 5	23	...	2334 2334 2335
	trivial atom number	2 3 4	34	...	323 434 433
E. Evict the following components from further calculation					
	rotatable atom number	5	0	...	0
	trivial atom number	1	0	...	0
F. The new 2D matrix					
	rotatable atom number	5 5 5 6	23 23 24 24	...	2334 2334 2335
	trivial atom number	2 3 4 8	34 55 59 1011	...	323 434 433

Figure 7

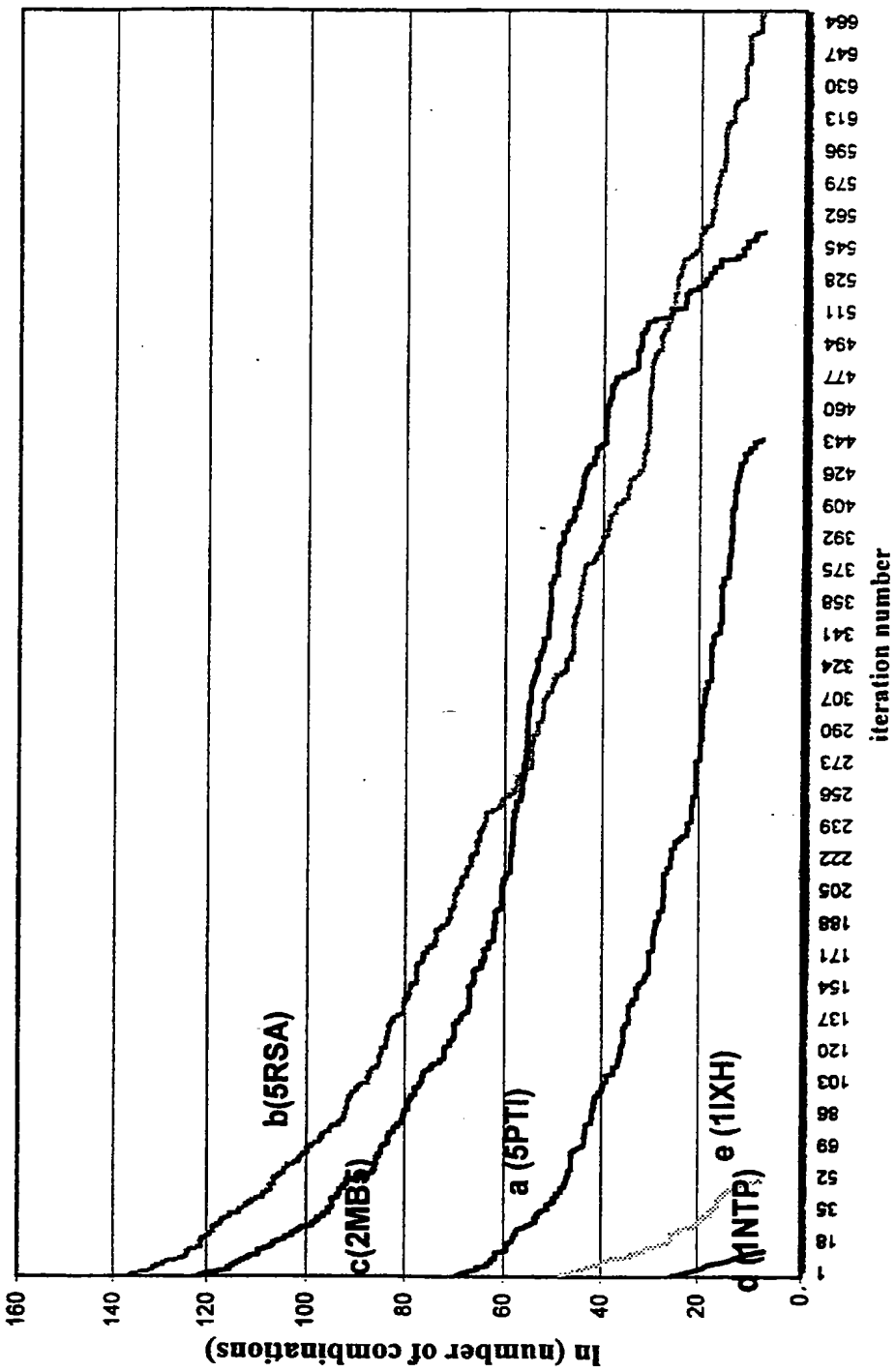


Figure 8

Figure 9

Figure 9A

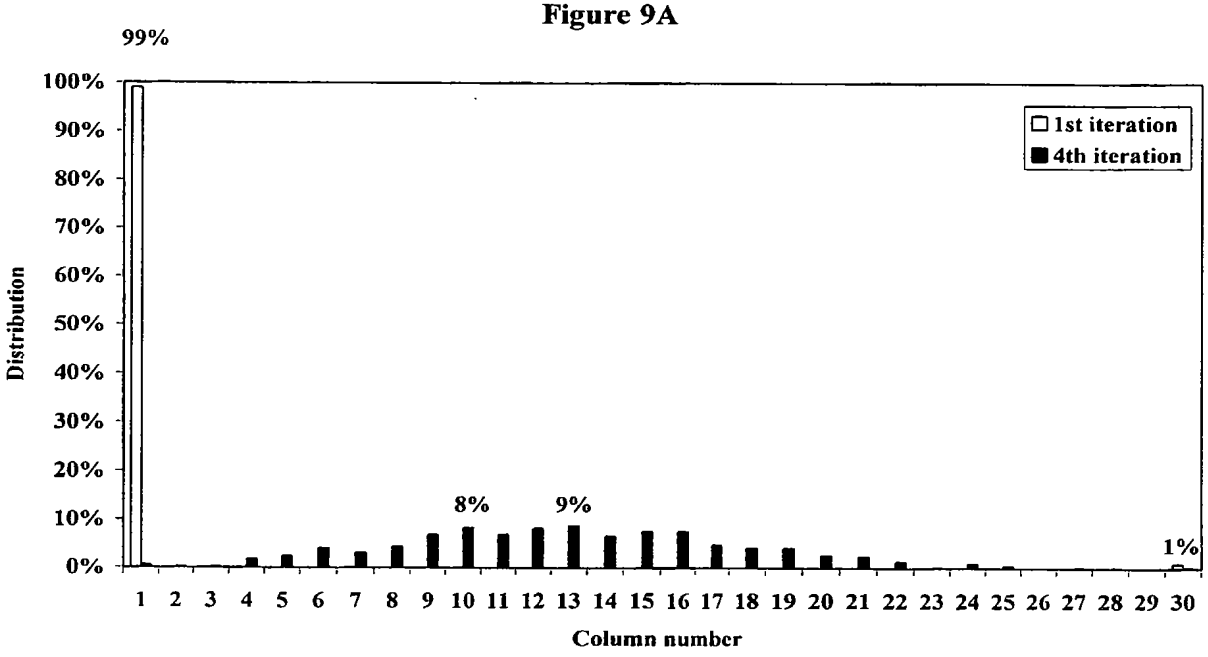


Figure 9B

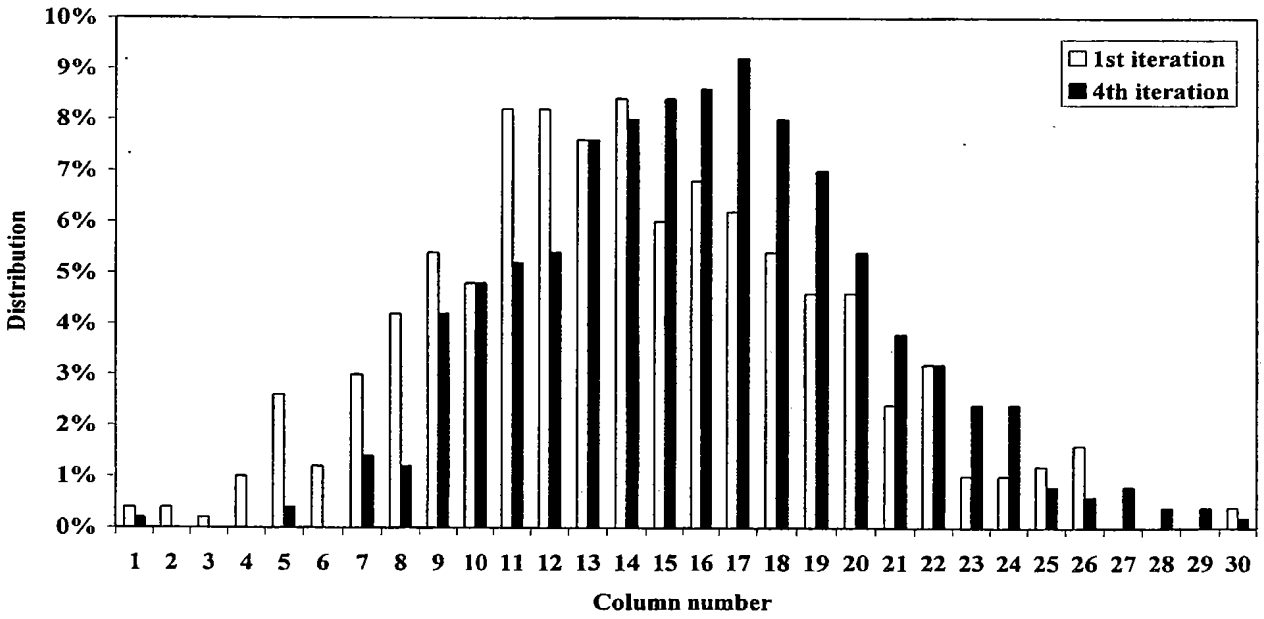


Figure 9

Figure 9C

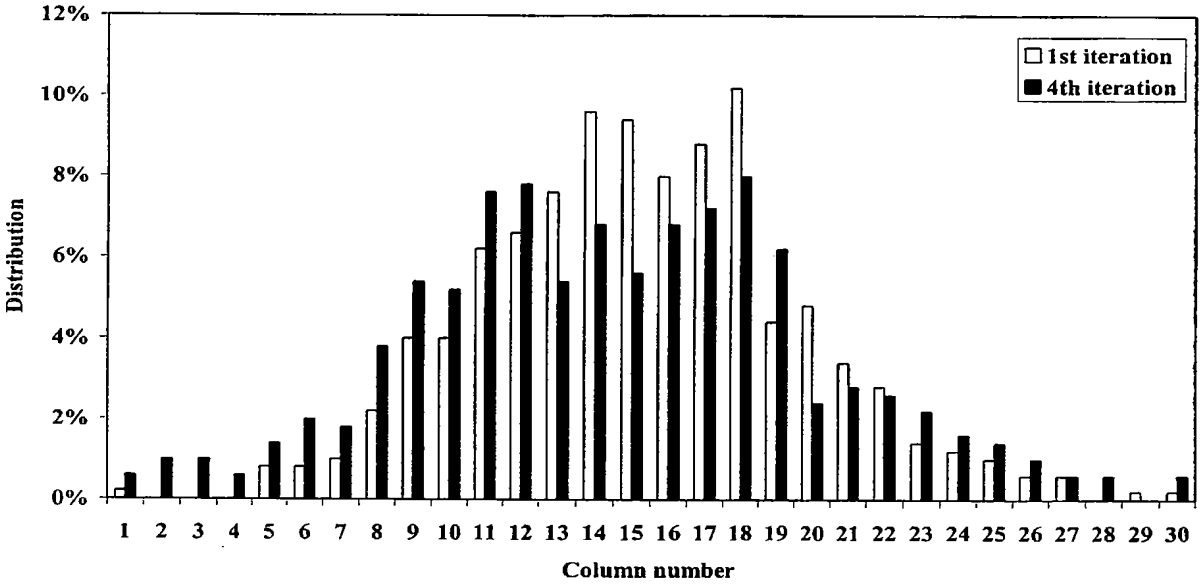


Figure 9D

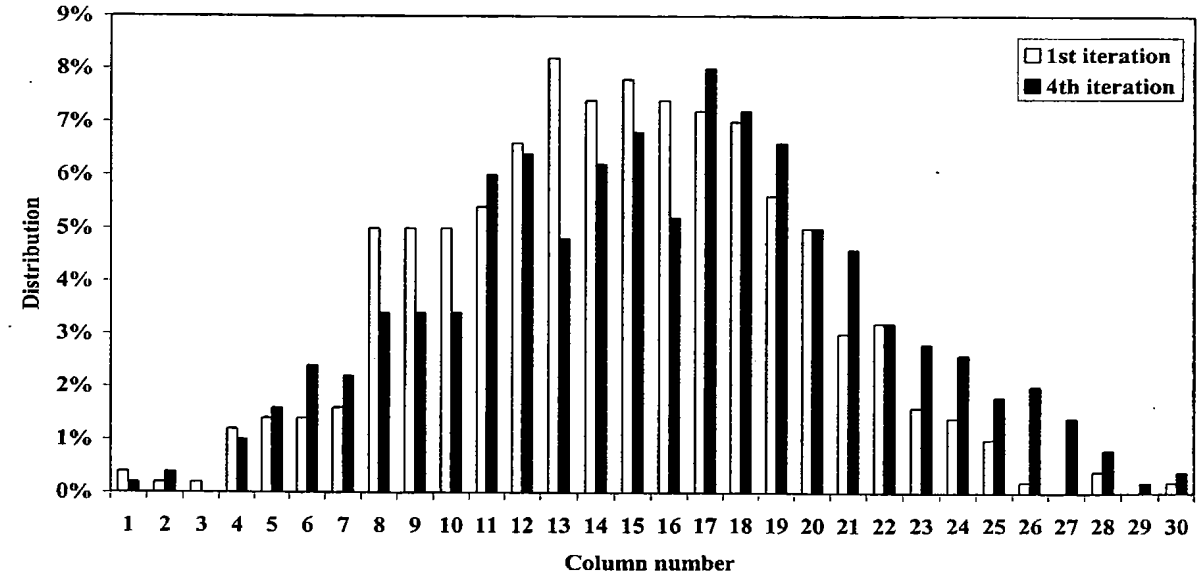


Figure 10

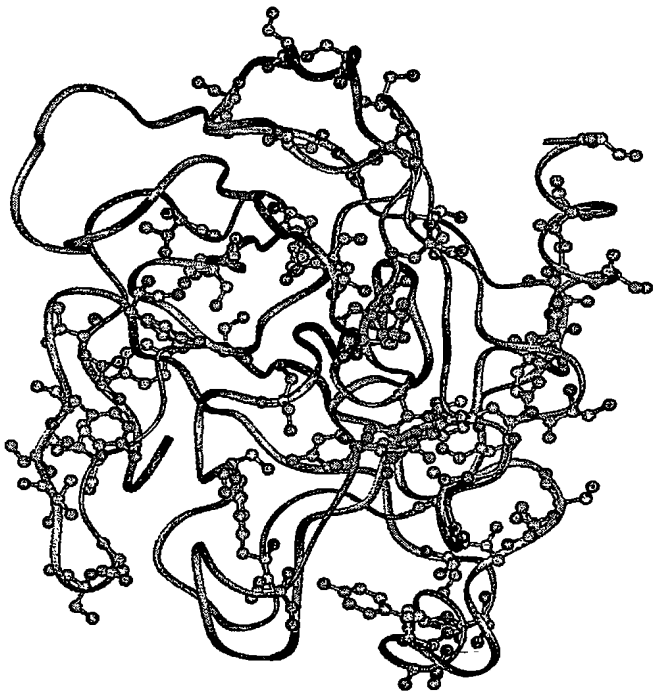
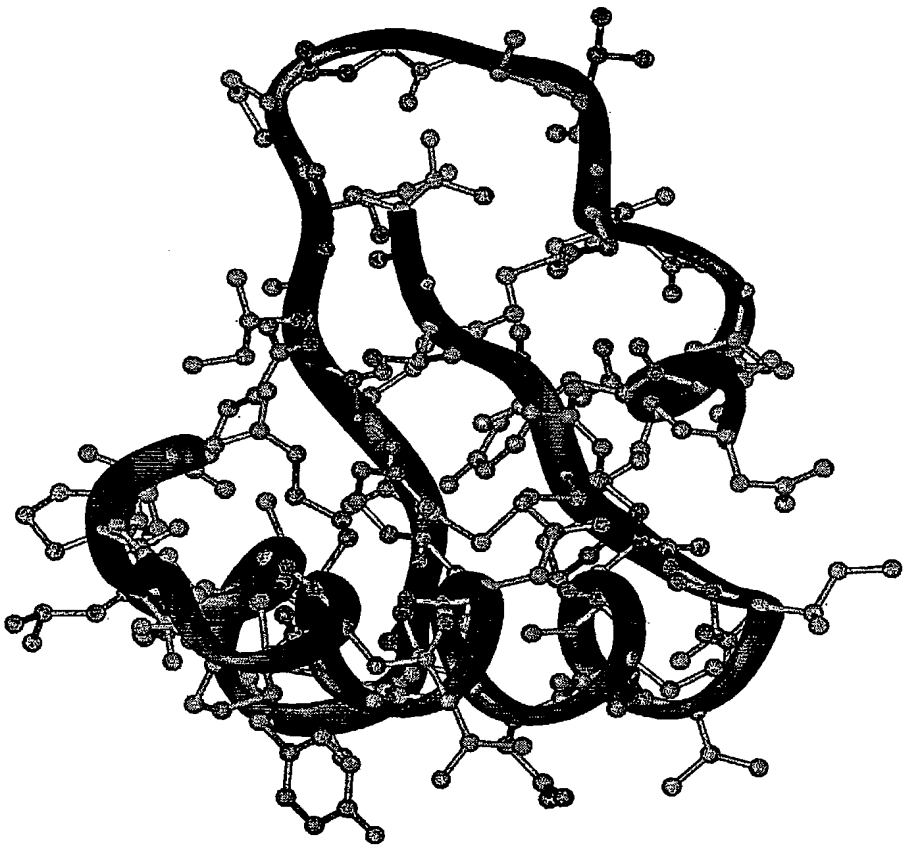


Figure 11



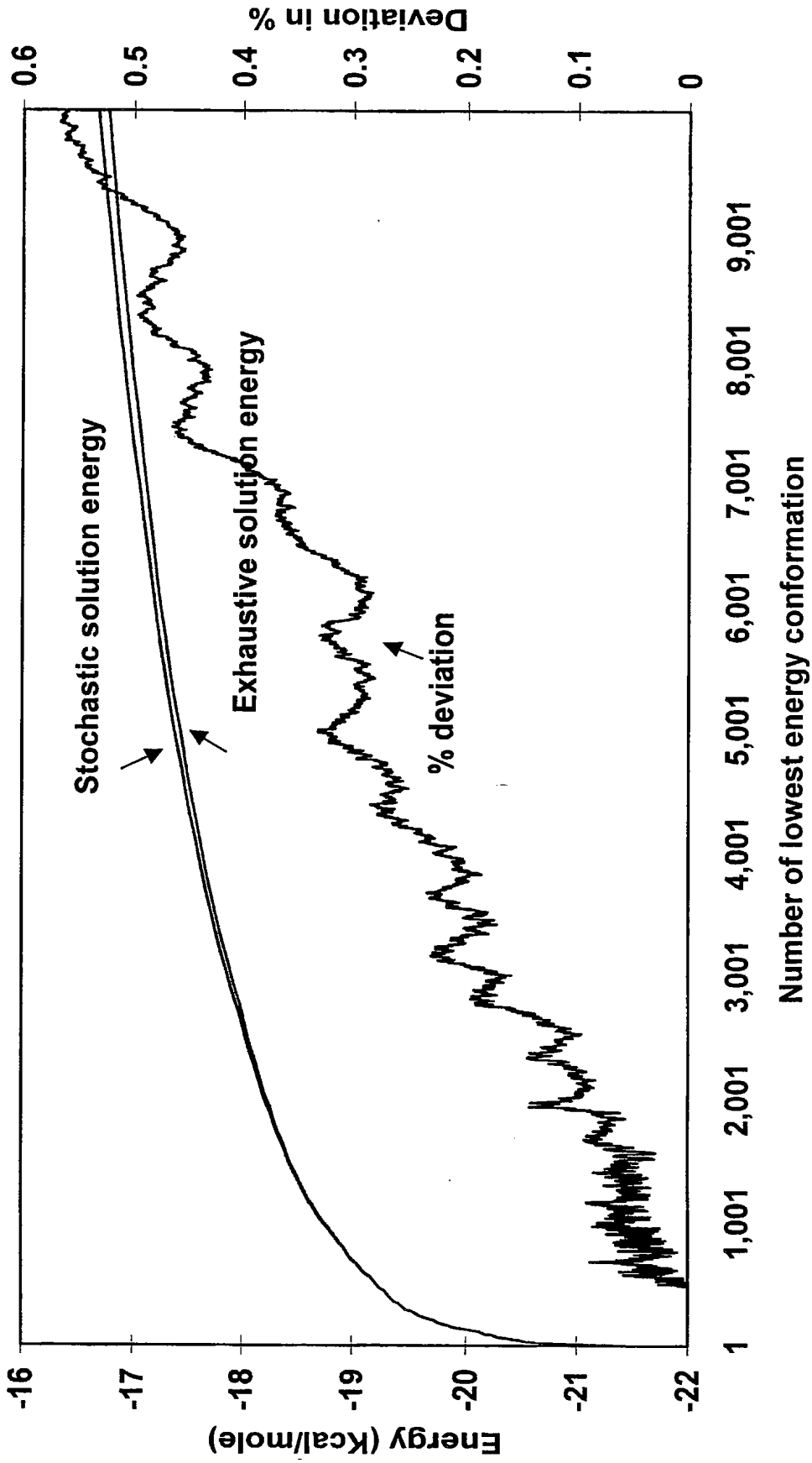


Figure 12

Figure 13

Figure 13A

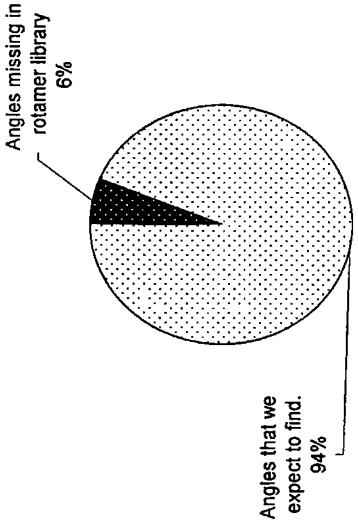
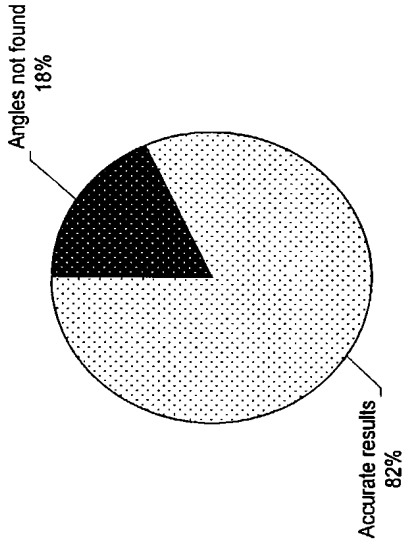


Figure 13B



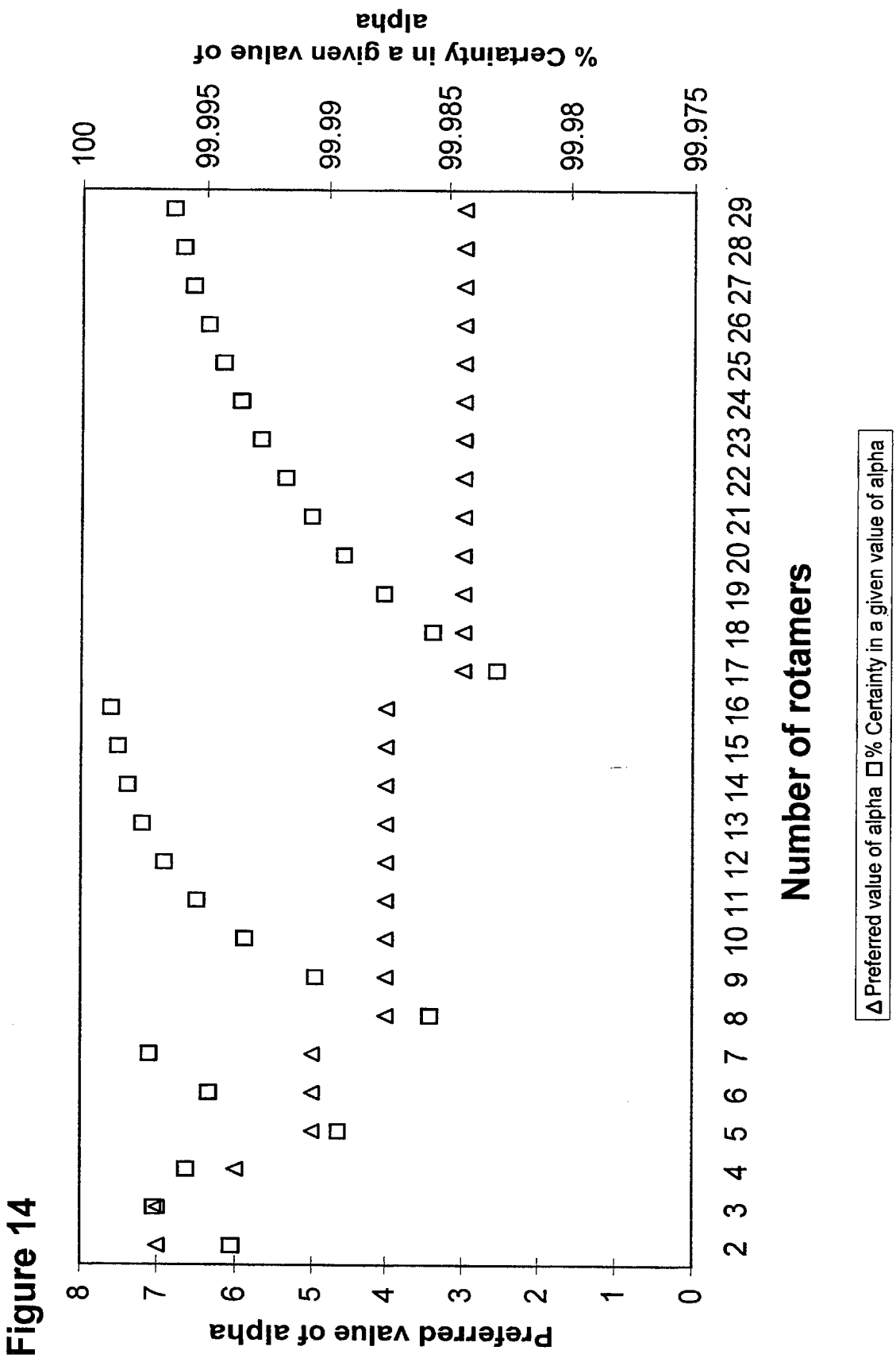


Figure 14

Figure 15

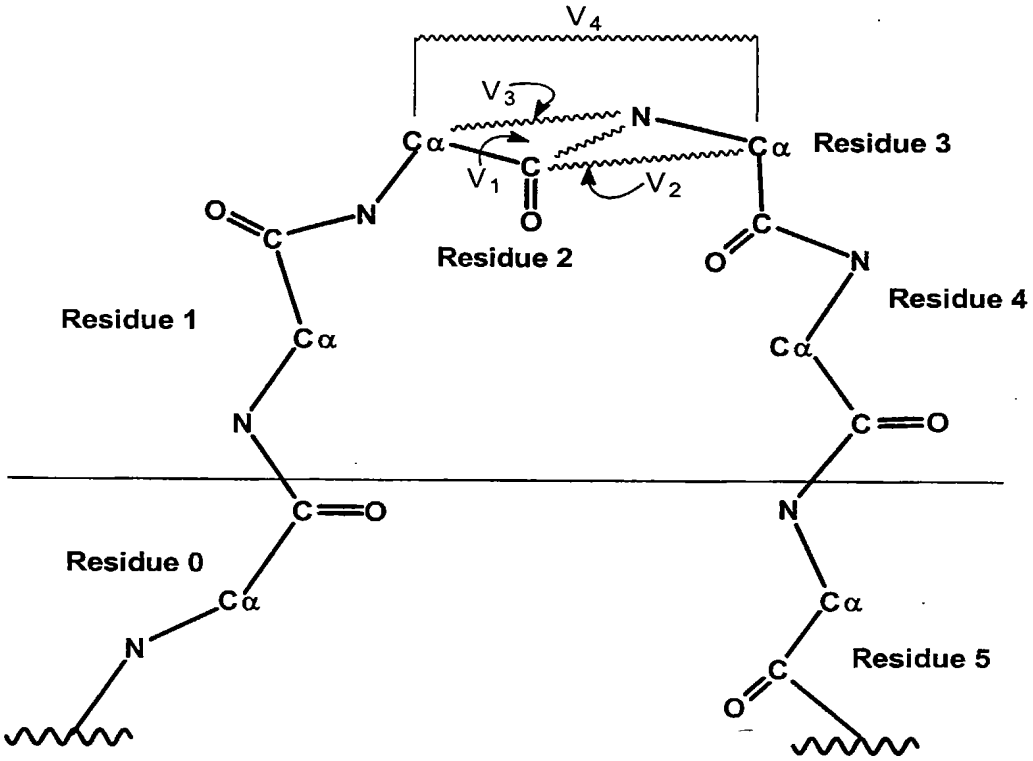
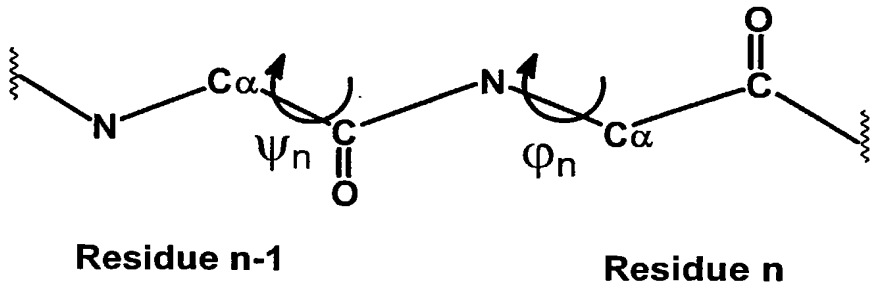


Figure 16



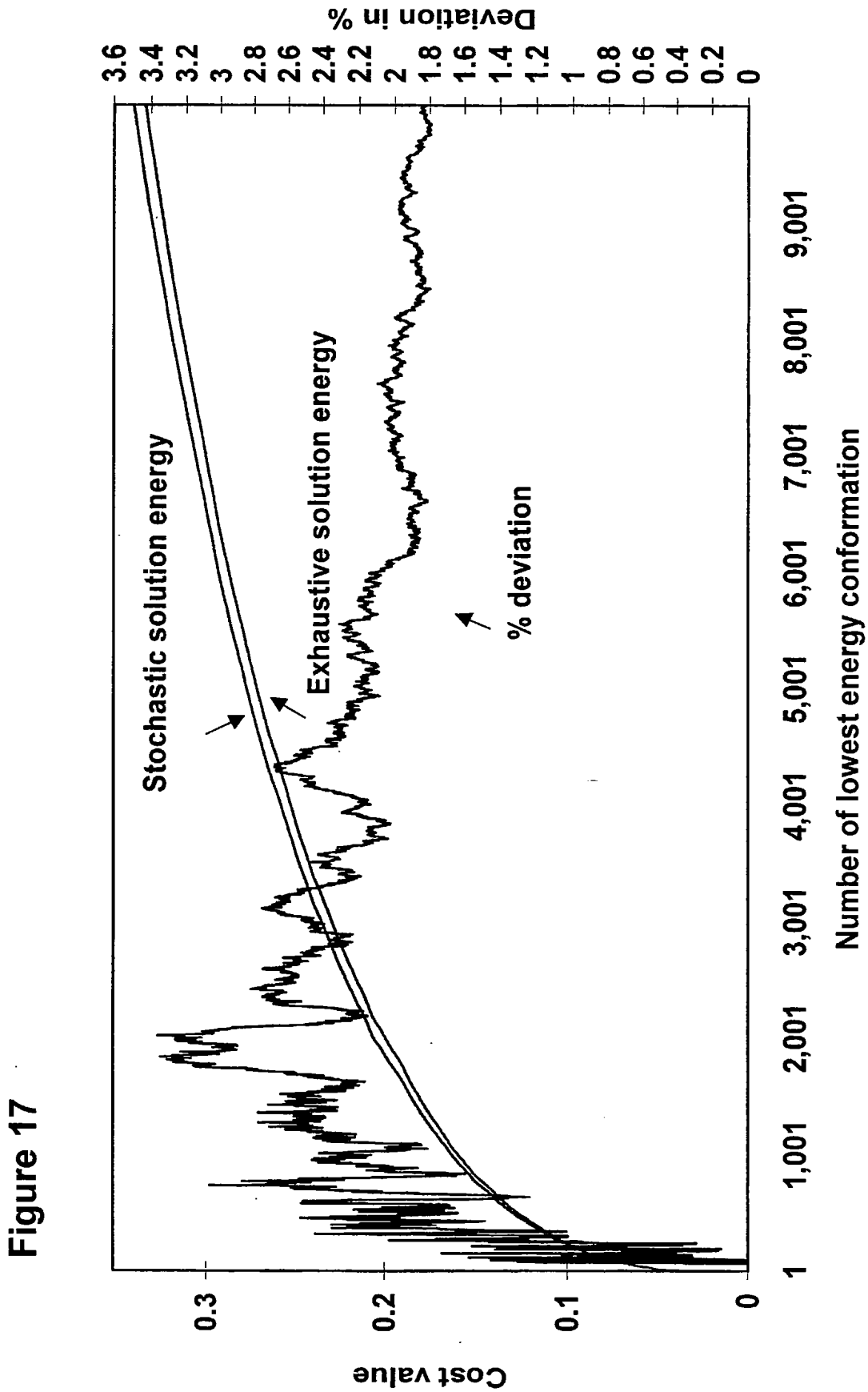


Figure 17